RIVERSIDE TEXTBOOKS
IN EDUCATION

EDITED BY ELLWOOD P. CUBBERLEY

DEAN EMERITUS, SCHOOL OF EDUCATION
LELAND STANFORD JUNIOR UNIVERSITY

STATISTICS FOR TEACHERS

BY

ERNEST W. TIEGS, Ph.D.

DEAN OF UNIVERSITY COLLEGE
UNIVERSITY OF SOUTHERN CALIFORNIA

AND

CLAUDE C. CRAWFORD, Ph.D.

PROFESSOR OF EDUCATION
UNIVERSITY OF SOUTHERN CALIFORNIA

HOUGHTON MIFFLIN COMPANY

BOSTON · NEW YORK · CHICAGO · DALLAS
ATLANTA · SAN FRANCISCO

The Riverside Press Cambridge

COPYRIGHT, 1930
BY ERNEST W. TIEGS AND CLAUDE C. CRAWFORD

ALL RIGHTS RESERVED INCLUDING THE RIGHT TO REPRODUCE
THIS BOOK OR PARTS THEREOF IN ANY FORM

The Riverside Press
CAMBRIDGE MASSACHUSETTS
PRINTED IN U.S.A.

EDITOR'S INTRODUCTION

ONE of the distinguishing characteristics of the great change which has taken place in educational work, within the past twenty years, has been the application of new statistical procedures to the study of educational problems. Whereas two decades ago we dealt with our problems in the light of accumulated experience, and by the aid of common sense and logical thinking tried to formulate the principles of action involved, to-day we attempt to determine principles from a study of concrete situations, and in turn to test the validity of the principles thus derived by a statistical treatment of the data from which the principles or conclusions were deduced. The concrete data for the study of the problem must be gathered accurately and after a carefully formulated plan, and the data must be organized and treated according to mathematical procedures. By such a plan we seek to eliminate errors, to reveal relationships, to clear up little-understood implications, to determine meanings and indicated trends which the data embody, and not only to deduce provable conclusions but in addition to be able to state the probability of error and its extent in the conclusions so deduced.

As a consequence of this change in the method of studying our educational problems, a new educational literature has been created, within recent years. New terms have been introduced into current educational thinking that are meaningful and need defining, new types of graphic display have been employed that need to be read intelligently, and new mathematical formulæ, which two decades ago were

almost unknown, have been derived and applied to the study of educational and social and psychological problems. As a result of this marked change in our methods of treating observational and recorded data, much of our educational literature of to-day is not intelligible to the reader who has no grasp of the newer methods for the statistical treatment of educational, social, and psychological data.

To meet the need for training in these newer procedures, courses in the statistical treatment of educational problems are now given in practically all schools of education, and the needs of the classroom teacher, who cannot take graduate courses in education, are being supplied through extension and summer-session courses in the elements of statistical methods as applied to the study of problems in educational practice.

It is this common need of teachers and beginning students of education that the authors of the present volume have tried to meet by preparing a textbook that will enable the teacher or student who studies it to follow intelligently the current educational literature. They have not tried to produce another comprehensive volume on statistical procedures as applied to education, or to education and psychology, but rather to present the elements of statistical science as applied to the one field of education, and to present these elements in such a simple manner that the beginner in the study, with little or no mathematical background, may grasp and be able to understand the fundamental principles of statistical procedures as now commonly employed in the educational literature of the day. Still more, the text has been constructed largely on a learning-by-doing plan, supplementing the usual statistical theory by adequate practical exercises, and with a view to preparing the teacher to apply simple statistical procedures in connection with the daily work of the school. These

purposes the authors have carried out very effectively, presenting in clear and simple language a volume which can be strongly recommended for classroom use as a beginning textbook in statistical methods for students of education.

Ellwood P. Cubberley

PREFACE

Any one who has the courage to add another volume to the long and growing list of textbooks in the field of statistics must necessarily do so for good cause. The authors believe that this work will justify itself, and that it will meet a definite need in the training of teachers and other school workers. Below are a few brief statements of what they consider the essential characteristics of the work:

1. It aims to meet the needs of the very large number of unspecialized educational workers, rather than the needs of the much smaller number of highly specialized educational research experts. The former group is typified by the classroom teacher, who has given the work its name. The research expert may use the book as a first step in his training, but will need to follow it with work in more advanced treatises.

2. It is addressed to educational workers only, rather than to those in such fields as sociology, economics, or business. While principles and processes are rather similar in all fields, the topics and illustrations have been selected specifically to meet the needs of teachers. It has seemed best to serve teachers well, rather than to seek a wider audience at the expense of more direct application of what is taught.

3. It is constructed for use as an actual classroom textbook or manual. It is intended to be the major, if not the only, book to be used in the course, and is arranged in the order which the authors find it best to teach the various steps and processes. The aim has not been to present a complete compendium of statistical information, but rather

to organize a set of definite lessons in convenient and sound instructional order.

4. It is provided with a set of true-false tests at the ends of the chapters for the use of students in studying, and for the use of the teacher in developing an understanding of the theory or principles which are included. If these tests are used effectively, they should go far toward cultivating statistical insight, or toward preventing statistical work from being mere mechanical juggling.

5. It attempts to develop actual skills, rather than merely to impart knowledge or information about statistical operations. For this reason numerous problems and exercises have been placed at the ends of chapters, as well as examples within the chapters. The authors recommend that students be required to do all the exercises, and that, if necessary, some classes be asked to do additional exercises based on data supplied by the instructor, in order that an adequate degree of skill may be attained.

6. It approaches new concepts, principles, or processes by way of concrete cases, whenever possible, and also applies these concepts or principles to concrete situations or actual educational problems after they have been explained and illustrated. Most of the model computations are based on actual rather than imaginary data, in order that learning may be carried on under conditions as similar as possible to real life situations.

7. It correlates graphics and statistics by presenting them both throughout the work, instead of including a chapter on charts and graphs as a separate unit. The authors think of graphs not as "statistical luxuries," but rather as "statistical necessities."

8. It seeks to avoid highly technical or theoretical phases of mathematics. Only the common elements of arithmetic and occasional phases of algebra are taken for granted.

In actual practice the authors have found it useless to assume more than this.

9. It does not attempt to derive formulas nor to explain their mathematical foundations if they are at all complex or difficult; the same amount of time probably can be spent to better advantage in a beginning course if devoted to learning to use formulas and to interpret results. The mathematical proofs of formulas are recognized as being valuable, but are omitted because of the impossibility of accomplishing all desirable objectives in a single elementary course.

10. It omits entirely, or else mentions very briefly, many processes and concepts which are often treated in statistics texts. This is done from necessity rather than preference, lest the effort to teach everything result in failure to teach well the minimum essentials.

11. It has been written as an outgrowth of the experience of the authors in teaching statistics to a considerable number of classes, as well as in directing scores of graduate thesis investigations, many of which were of a statistical nature. The selection and the presentation of topics has been influenced by what seemed to be the real needs and the real difficulties in practical situations.

12. It has been tried out in mimeographed form with four classes, taught by three different instructors and has been revised in the light of these practical tests.

13. It has been read and criticized very constructively by Dr. Walter E. Eells, of Stanford University. A number of very important revisions and improvements have resulted from these criticisms.

The authors are indebted to a number of persons for helpful suggestions, criticisms, or special courtesies shown in the preparation of the manuscript. Among these are the following: (1) Dr. W. S. Ford and Dr. O. R. Hull, Professors

of Education in the University of Southern California, for numbers of practical suggestions. (2) Mr. C. H. Nettels, Assistant Director of the Division of Psychology and Educational Research, of the Los Angeles City Schools, for helpful comments growing out of his experience in teaching the mimeographed edition of the book in his summer school classes at the University of Southern California. (3) Dr. Walter C. Eells, Professor of Education at Stanford University, for courtesies already mentioned. (4) Dr. Ellwood P. Cubberley, Dean of the School of Education at Stanford University, for editorial criticisms and improvements in the manuscript. (5) The students who were in the classes taught by the writers and by Mr. Nettels, both before and after the writing of the book, for their frank reactions as well as encouraging interest in the work.

ERNEST W. TIEGS
CLAUDE C. CRAWFORD

UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES

CONTENTS

I. Why Study Statistics 1
II. Labor-Saving Devices and Equipment . . 12
III. Tabulating Data 28
IV. The Mean 49
V. The Median 62
VI. Percentiles 73
VII. Variability 91
VIII. The Normal Curve 106
IX. Reliability 129
X. Correlation 153
XI. Further Aspects of Correlation . . . 174
XII. Principles of Statistical Research . . 191
Selected References 205
Index 207

LIST OF TABLES

I. Illustration of a Simple Frequency Distribution, with Tabulation 29
II. Illustration of a Frequency Table Grouped in Class Intervals 31
III. Illustrations of Methods of Designating Class Intervals 36
IV. Illustration of the Method of Finding the Mean from a Frequency Table, Using Zero as the Starting-Point . 52
V. Illustration of the Method of Finding the Mean when the Intervals are Larger than One, and Using an Assumed Mean 54
VI. Illustration of the Method of Checking the Accuracy of the Computation of a Mean by Using a Different Assumed Mean 56
VII. Illustration of the Method of Finding the Median of Data Grouped into a Frequency Table 64
VIII. Illustration of the Cumulative Frequency Table . . 77
IX. Illustration of the Process of Converting Cumulative Frequencies into Percentile Ranks of the Tops of Class Intervals 78
X. Illustration of the Method of Finding the Percentile Ranks of the Midpoints of Class Intervals . . . 80
XI. Illustration of the Computation of Percentiles for Ungrouped Data 88
XII. Illustration of the Computation of Quartile Deviation 97
XIII. Illustration of the Computation of the Standard Deviation 99
XIV. Illustration of the Computation of Quartile and Standard Deviation 101
XV. Illustration of the Possible Combinations of Heads and Tails when Two Coins are Tossed 107
XVI. Illustration of the Possible Combinations of Heads and Tails when Three Coins are Tossed 108

XVII. Percentages of the Normal Distribution Lying Between the Mean and Various σ Unit Values . . . 115
XVIII. Percentages of Students Assigned Each Mark Under Various Systems that Use the Normal Curve as a Basis for Marking 117
XIX. Chances that Measures will not Fall Below Various ϵ Unit Values 137
XX. Illustration of the Computation of Various Quantities as they are Involved in a Rotation Experiment 143
XXI. Chances that Measures will not Fall Below Various *P.E.* Unit Values 148
XXII. A Scatter Diagram Showing Correlation of .79 . 157
XXIII. Illustration of the Pearson Product-Moment Method of Computing the Coefficient of Correlation from a Correlation Table . . . *Between* 158 *and* 159
XXIV. Illustration of the Computation of the Pearson Product-Moment Correlation Coefficient from Ungrouped Pairs of Data 165
XXV. Illustration of the Computation of the Coefficient of Correlation by the Spearman Rank Differences Method 167
XXVI. Data to be Used for Practice in Computing Coefficients of Correlation 172
XXVII. Data for Practice in Connection with Special Aspects of Correlation 190

LIST OF FIGURES

1. Graphic Illustration of a Class Interval 33
2. Illustration of the Graphic Representation of a Frequency Distribution: Frequency Rectangles 38
3. Illustration of the Graphic Representation of a Frequency Distribution: Histogram and Frequency Polygon . . 40
4. Illustration of the Comparison of Groups by Means of Frequency Polygons 41
5. Illustration of the Difficulty Involved in Comparing Groups of Unequal Size by Plotting Simple Frequencies 42
6. Illustration of the Comparison of Unequal Groups by Plotting Percentage Frequencies 44
7. Illustration of the Construction of a Percentile Graph . 83
8. Illustration of the Use of Percentile Graphs to Compare Two Groups 86
9. Illustration of the Need of Taking Account of Variability when Comparing Two Distributions 93
10. Illustration of the Curve Obtained from Plotting the Expanded Binomial 110
11. Illustration of the Percentages of the Total Distribution Corresponding to Successive σ Units from the Mean 113
12. Illustration of the Relation Between σ Unit Values and Fractional Parts of a Normal Distribution 116
13. Illustration of the Method of Dividing the Normal Distribution Curve into Five Parts for the Assignment of Teachers' Marks 119
14. Illustration of the Computation of Difficulty Values for Five Test Questions 121
15. Illustration of the Basis for Converting an Observed Difference into a Statement of Chances that the True Difference is Above Zero 141
16. Illustration of the Steps in Finding a Partial Correlation Coefficient 181
17. Illustration of the Steps in Finding a Multiple Correlation Coefficient 183
18. Illustration of the Meaning of Curvilinear Correlation . 187

STATISTICS FOR TEACHERS

CHAPTER I

WHY STUDY STATISTICS?

Need for a knowledge of statistical procedures. The stories of scientists who became so absorbed in their researches that they forgot their dinners or their weddings have their counterparts in the lives of statisticians. Statistical work is extremely fascinating. Figures are not dry! Under the direction of a trained statistician, statistical procedures can perform feats quite as remarkable as the tricks of a Houdini. The solution of statistical problems yields a satisfaction as deep and as natural as that which a veteran chess player gets from directing his campaign, or an inventor derives from the creation of a new machine.

And yet the major reason for studying statistics is not that of interest. The most vital reason is utility, or practical value. A teacher must study statistics, because without it he cannot hope to get far in education, or even be a highly successful teacher in this generation.

This need for statistics on the part of educational workers manifests itself in three separate areas of their educational responsibilities and activities. A knowledge of statistics is necessary in order to understand and interpret educational science, in order to do the day's work most effectively,

and in order to contribute to our present knowledge about education. We shall consider each of these more fully.

Need for statistics in the study of educational science. The elementary concepts of statistics constitute a part of the common culture or heritage which integrates the teaching profession. A teacher who is not familiar with means, medians, percentile ranks, standard deviations, or correlations is in a situation comparable to that of the person who attends religious services conducted in a foreign tongue. He may watch his neighbors and imitate them, but he neither understands, participates, nor contributes in an intelligent or effective manner.

A considerable body of educational literature, increasingly more statistical in nature, has been published in recent years, and a growing number of new books and journals of this type are published every year. Some years ago, the writers of educational books and articles apologized to their readers if they were forced to use statistical terms, even of the most elementary kind; to-day they take a knowledge of statistics for granted and do not hesitate to publish material of a highly technical nature. The improved statistical training of the readers has made this possible, but at the same time this tendency on the part of educational writers has created the need for more adequate statistical training on the part of readers. Teachers to-day will be able to read professional literature intelligently and with profit in proportion as they gain understanding in this field.

A knowledge of statistics is essential to the teacher in order to use intelligently the findings of the educational

research experts. Profound theses embodying valuable work for the Ph.D. degree are often "buried" because those who should apply the findings in the educational activities of the schools cannot understand the research reports. If we are ever to succeed in bridging the chasm between the research experts and the practical educational workers, we shall have to build from both sides of the gulf which now separates them. The study of statistics by practical school people will enable them to meet the research workers halfway.

Statistical training good protection against errors. Education, like medicine, law, and politics, is infested with many "isms" and "ologies." We have numerous quacks, shysters, and false prophets, each recommending his pet theory, hobby, or scheme, and enlisting teachers in carrying out his program. With so many available leaders, whom shall the teacher follow? Our educational democracy desires increasing freedom for each teacher in making decisions and plans; this freedom must, therefore, be accompanied by a type of training which will enable him to use this freedom wisely. Statistical training is a good protection against the false and erroneous conclusions of others. It is said that figures can't lie but that liars can figure. A propagandist can present absolute falsehoods supported by misleading statistical evidence and make them appear to be the essence of truth — unless his audience knows enough about statistics to discover his tricks and statistical fallacies.

Need for statistics in the conduct of educational activities. A knowledge of statistics is useful in the daily work

of the teacher and the school administrator. Before school begins in the autumn there is need for an analysis of the available school-census data. Students have to be classified into grades, levels, and sections for instructional purposes, on the basis of quantitative data. The success of the year's work is very greatly increased if this phase of school organization and administration is carried out according to sound statistical principles, and by the aid of suitable statistical techniques.

Nor is this need for statistics confined to a limited few in the administrative offices. The classroom teachers usually bear a considerable burden in getting the school well started. One year's plans are made up in the light of the previous year's records, which are often prepared largely by classroom teachers. The task of organizing and administering a school is so great, and the interrelations between the various members of the school staff are so intricate, that each member, however insignificant he may seem to be, needs to know something about the larger scheme of things in order to coöperate more intelligently. An ignorance of statistical procedures on the part of the teachers tends to tie the hands of the best administrative staff, or the best school research department, just as an untrained army handicaps the best of generals.

Statistical procedures are valuable to the classroom teacher in handling daily routine. Teaching seeks to produce change in pupils, and this in turn calls for measurement of pupil achievement. Marks and grades occupy a very prominent place in the teacher's work, and without a

knowledge of statistics a teacher cannot meet his responsibilities intelligently. For example, a teacher without statistical training attempted to apply such knowledge as had been picked up in the distribution of pupils' marks. Six tests were given during the course, and the grades were distributed each time according to the normal curve. At the end of the course the teacher had no A grades, no failures, and over 50 per cent of the total were C grades. Although surprised, perplexed, and somewhat humiliated, the teacher did not know what was wrong with the method. Adequate acquaintance with the normal-curve hypothesis would have solved the problem easily.

Statistics a method of thinking. Often the teacher wishes to measure the effectiveness of some particular teaching procedures, and would like to know which of two ways of teaching yields the better results. If an experiment is conducted to answer the question, how can sound conclusions be drawn without a knowledge of the statistical procedures for analyzing his experimental data?

The classroom teacher is able to think more intelligently about teaching procedures if possessed of statistical training. Statistics is a universal method of thinking rather than a mere specific tool. A person who has had adequate training in this field is likely to be more critical and scientific in thinking about ordinary non-statistical subjects. The teacher is less likely to fall a victim to emotions or prejudices in deciding issues involving teaching procedures if the importance of objectivity as it is exemplified in statistics has once been realized.

School officials find statistics invaluable as a means of gaining coöperation and support from the public. Statistical evidence is more convincing than mere statements of opinion. The public desires and has a right to know what the schools are doing, and with what success. An administrator who wishes to render an accounting to the community, or a classroom teacher who wishes to give the superior officer evidence of the success of the work being done, can do so with greater effectiveness if he knows how to collect, analyze, interpret, and present statistical data.

Need for statistics in the improvement of educational science. Educational research is complex and highly specialized. Statistical procedures bear the same relation to research that tools bear to carpentry. A mere "saw-and-hammer man" will not accomplish much in educational research. The problems that confront teachers are intricate and difficult. The procedures for solving them are often equally elaborate and technical. A problem that seems hopeless to one teacher may be solved easily by another who has the proper techniques and skills with which to attack it. Every teacher must to a certain extent be a research worker, because if teachers do not solve some of their own professional problems many of them are likely to go unsolved.

Statistics lends precision to research. The day has passed when educational science rested on a basis of adjectives; the day of numbers is at hand.

Statistics is particularly essential in education and the social sciences, and, in fact, in all those fields of research

where individual differences are noticeable. If all children reacted alike, there would be much less need for statistics. We could observe one child and apply our conclusions to all children. In physics and chemistry we find a great deal of such uniformity of reaction, and a correspondingly small emphasis on statistical method. In education, however, the existence of individual differences makes it necessary for us to observe large numbers of cases in order to determine general trends that characterize them. Statistics constitutes a very effective aid in making such generalizations. By means of it we can arrive at relatively universal truths in spite of the difficulties created by individual differences.

Statistics and the social sciences. Statistics is a valuable supplement to the experimental type of research. Many educational problems are difficult to attack by the experimental method. In the natural sciences the experimenter can control physical objects or lower animals, even to the extent of carrying on experiments with the aid of vivisection. We have no such freedom in education, and, if we did have, we should encounter another difficulty. The long infancy, or slow rate of maturity, of human beings forces us to wait too long for the results of some of our educational experiments. By the use of statistics, however, we may take children or adults as they are, collect information regarding their previous experience or educational treatment, and draw conclusions comparable to those which would result if we were able to control all these factors experimentally. In other words, statistics is a method by

which we summarize the outcomes of a large number of individual experiments which nature or environment has already performed without our control or influence.

Statistics bears somewhat the same relation to educational and social research as high school and college mathematics bears to research in the natural sciences. The mathematics courses with which we are familiar were evolved in an earlier day when the physical sciences were in their infancy, and before quantitative methods had begun to be applied to social problems. Society has progressed far since then, and a need for a different type of mathematical training has been created. In the same manner that the traditional type of mathematics has served the physical sciences and technical studies, statistics is essential in the solution of social problems.

There is a growing tendency to substitute statistics for at least a part of the foreign languages that usually are required for the Ph.D. degree. A few decades ago educational research was chiefly historical in nature, and foreign languages were needed more or less in order to provide access to essential source materials. Furthermore, educational thought was further advanced in Europe than in America, and there was greater need than at present for a reading knowledge of French and German. The shift of emphasis from the past to the present, and the shift of the educational research center of gravity from Europe to America, have both increased the demand for statistics as a tool of research and reduced the need for ability to read French and German. This shift of emphasis is manifested

in a number of graduate schools of education by the acceptance of an elementary course in statistics in place of one foreign language, and in some schools by the acceptance of advanced statistics in the place of the second foreign language.

TRUE–FALSE TEST

NOTE: *The exercise at the end of each chapter labeled " True-False Test" is not an examination in the usual sense. Its primary purpose is self-testing by the student as he prepares his assignment. In the ordinary objective test, the items should admit of only one correct answer; but there are many important considerations which the student should master about which no such unequivocal statements can be made. Throwing them into this form challenges the student to find and weigh the evidence upon which an answer depends. For measurements, the instructor may select the items which lend themselves to his purpose.*

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. Statistical procedures enable us to generalize from only a few cases.
2. Statistical procedures are especially important in the natural sciences.
3. The study of statistics leads to confusion in thinking about non-statistical subjects.
4. The average citizen has practically no need for training in statistics.
5. The study of statistics aids one to see the weaknesses of another person's work.
6. The general public is more prejudiced against statistical reports than against non-statistical ones.
7. Statistics constitutes almost the total of research technique needed by the scientist.
8. The knowledge of statistics acts as an integrating force for progressive educators.

9. A knowledge of statistics makes possible the replacing of numerals by meaningful adjectives.
10. There is a fundamental difference in point of view between the statistical and the experimental types of research.
11. The reason why statistics is usually thought of as dry is that it is so difficult to learn.
12. Statistics is simply a special type of thinking.
13. Pages of statistical data in support of contentions make boards of education suspicious of the facts concealed.
14. The authors recommend that teachers always apply the normal curve in distributing marks.
15. It is impossible for prejudice and bias to get a foothold in a statistical investigation.
16. Educational measurement is of relatively small value without statistics.
17. The creation of test and measurement bureaus in public schools reduces the need for statistics on the part of the teacher.
18. The principal reason for studying statistics should be the development of ability as a research worker.
19. Language requirements for the Ph.D. degree should be abandoned.
20. Courses in statistics should replace those in high-school mathematics.

QUESTIONS AND EXERCISES

1. Why is the study of statistics ordinarily thought of as being dry?
2. What new activities in the daily work of the classroom teacher call for statistical training which was not needed twenty years ago?
3. What portion of the recent advance in the science of education would you ascribe to statistics as compared with the other factors or forces?

4. Explain how statistics is a tool for generalization.
5. What percentage of the faulty conclusions drawn from statistics would you say are intended to be misleading?
6. Compare the *Journal of Educational Psychology*, the *School Review*, *School and Society*, the *Journal of the National Education Association*, and the *Journal of Educational Research*, as regards the amount of statistical training required of the reader.
7. Why are master's theses usually "buried"?
8. What are the differences between the experimental and the statistical viewpoints?
9. What, in your opinion, is the relative value of a course in statistics and a reading knowledge of French or German as a preparation for research?
10. What would be the gain and what the loss if a course in statistics were substituted for high-school algebra? For high-school geometry? For trigonometry?

CHAPTER II

LABOR-SAVING DEVICES AND EQUIPMENT

Importance of labor-saving devices. Expert statisticians employ numbers of short cuts and labor-saving devices to speed up their work and reduce the routine involved in making calculations. Some of these short cuts are in the nature of procedures, and others may be classified as forms of equipment. Throughout this book we shall be dealing from time to time with the special methods of shortening the work in the different statistical processes, but it is our aim in the present chapter to make an acquaintance with a few of the principal types of equipment which those who do statistical work on a small scale may find serviceable. A student who expects ever to do any statistical work should purchase, at the beginning of his course, some of the simpler and less expensive labor-saving devices which he will use during the course and afterward. For a student to fail to secure any such equipment for the statistics course is simply a confession that he never expects to use statistics after he completes his training.

It should be understood at the beginning that any short cut or labor-saving device is likely to be more difficult to operate, and thus require more time at first than the longer and more laborious methods which are already known. The long method, however, fails to improve appreciably with practice, while the short method is likely to manifest considerable gain and soon prove far more economical than the

long one. Also, work done by short-cut methods is likely to suffer temporarily as regards accuracy, and may well be checked, during the early stages while accuracy is being acquired, by comparison with the long methods of calculations.

The following labor-saving devices are commonly employed by statisticians and merit the careful consideration of the beginning student.

Tables of squares, square roots, products, quotients, etc. Many statistics books include, at the back, tables showing the squares, square roots, reciprocals, etc., of the numbers most commonly used in statistical work. Tables also are available showing the products and quotients of numbers, so that the student may eliminate most of the multiplying and dividing involved in statistical work by simply turning to the desired page and reading the answer which has been computed by some one else. Similarly, it is possible to avoid a great deal of calculation of purely statistical quantities, such, for example, as the probable error of a correlation coefficient, by turning to a table in which the probable errors for correlations of all sizes and for various numbers of cases have already been computed and published. Some statistics books include tables of all these types at the back. Usually, however, the tables so included tend to be somewhat abridged and cover only a limited range. It is possible, however, to purchase a regular book of tables which will serve ordinary purposes and which may be used in connection with any regular statistics text.

Interpolation. One difficulty which is commonly found

in using tables is that they are often based on whole numbers below 100 and the student is sometimes obliged to get the products, quotients, or square roots of numbers involving fractions or numbers over 100. Thus, the table may show the product of 38 times 73 and also 39 times 73, but the particular problem may call for the product of 38.5 times 73. In such a case the student must find his desired product by taking a number half-way between the product of the two whole numbers in question or if he desires the product of 385 × 73 he must do the same as for 38.5 × 73 and set the decimal over.

This process is called interpolation. It may be employed to fill the gaps in any ordinary table, once the principle of it has been learned. The essential point as regards the method of interpolating is to find the quantity next above and below the desired one, get the difference between the two, and add to the lower number that fraction of the difference which corresponds with the decimal in the original problem. Thus, the product of 38.2 times 73 is found by getting the product first of 38 times 73, which is 2774, then getting the product of 39 times 73, which is 2847, finding the difference between these which is 73, and adding .2 of this difference to 2774, yielding 2788.6.

Another difficulty closely similar in nature to the one just mentioned results from the fact that one may have to deal with numbers which are smaller or larger than those given in the tables. Thus, a table of squares and square roots may run only up to 1000 and it is desired to find the square root of the number 1500. The solution of this diffi-

culty is found by locating the number 15 in the table, finding its square root, which is approximately 3.88, and then setting it over one decimal making 38.8, the approximate square root of the original number 1500. Similarly, if the problem is to divide 170 by 21 and the table being used only gives quotients up to 100 divided by 99, the student may find the quotient of 17 divided by 21, and then set the decimal place over to compensate for the change. In multiplying or squaring a number which goes beyond the table the problem is slightly different, and it is necessary to add ciphers to the product, thus securing answers only in round numbers. This is ordinarily permissible because if numbers are extremely large the last few digits are of relatively small importance anyway.

When using tables which carry decimals out to several places, it is frequently preferable to ignore the last few decimals in order to shorten work and make for economy, but in doing so it is very important that one more decimal be observed than is to be retained when the number is copied down. For example, if the number is .346 and we desire only to record the first two places, we notice the 6 in the third place and therefore write down our number .35; whereas, if the number in the table had been .344, we would have written it down as simply .34. In other words, we never know whether to keep the last figure as it is or raise it to one higher unless we observe whether the following digit is greater than 5. If the last digit is exactly 5 it would be as accurate to throw it away as to raise the number to the next highest whole number, but if one of these practices is

followed all the time there is a constant tendency to yield results either too high or too low. If a number like 34.5 is treated as 34 one time and 35 the next, the result in the end is more nearly correct. It has been suggested that the best rule for this problem is that of the Nautical Almanac and Coast and Geodetic Survey, "keep the last figure even." Thus 35.5 would be treated as 36, but 34.5 would be treated as 34.

Books of tables. The difficulties that have just been mentioned may be avoided by purchasing more elaborate tables, but doing so creates a new difficulty, because the tables become so bulky and complex that the physical task of finding the desired quantities is very great and the time required to look up an item is increased. Furthermore, most of the ordinary statistical work can be done with the abridged type of tables without difficulty. The problem is somewhat comparable to that of choosing between a handy dictionary that can be carried around and used conveniently, as contrasted with an unabridged dictionary which gives much more complete information at the expense of convenience.

The authors would recommend as the most valuable books of tables for education students:

1. Karl J. Holzinger's *Statistical Tables for Students in Education and Psychology*, published by the University of Chicago Press, 1925. This is decidedly the most convenient and well-selected set of tables available for general purposes, and includes not only tables of squares, square roots, products, quotients, and logarithms, but also various

statistical quantities, such as probable errors of correlation coefficients and areas and ordinates of the normal probability curve.

2. Barlow's *Tables of Squares, Cubes, Square Roots, Cube Roots, and Reciprocals of numbers up to 10,000,* published by Spon and Chamberlain, New York, 1924. This is a very complete set of tables for the particular items that are included, and the book has been widely used for a number of decades.

3. *Circumferences and Areas of Circles, Squares, Cubes, Square Roots, Cube Roots, and Reciprocals,* published by The Lefax Company, Inc., Philadelphia. This is a very inexpensive pocket-sized pamphlet, and contains numbers from .10 up to 1019.

Logarithmic tables. The use of logarithms for statistical calculations is quite common. Most books of statistical tables include logarithmic tables. By means of logarithms it is possible to eliminate practically all multiplication, division, and extraction of square roots. Thus, in order to multiply two numbers one finds the logarithm of each in the table, adds the two logarithms, and then finds the number which corresponds to the sum of these "logs." To square a number one finds its log, multiplies it by two, and looks up the number which corresponds with this product. Similarly, to get the square root of a number the log is divided by two. It is common for logarithmic tables to be carried only to four or five decimal places, but this yields a sufficient degree of accuracy for all ordinary statistical purposes.

If the student is already skillful in the use of logarithms

he will find them extremely valuable as an aid to statistical work. If, however, the whole subject is entirely new to him, the authors of this book are inclined to question the wisdom of investing the time required to learn how to use them. The same amount of time might be worth more if devoted to learning how to use such tables as were mentioned in the last topic, or in learning how to use the slide rule, which will be described next.

The slide rule. Many things that can be done by means of logarithms can be done also with the slide rule. In fact, the slide rule is simply a special device which applies the same principle as logarithms. It may be described as a mechanical device in which the logarithms of numbers are represented by distances on the rule, and which has a movable slide which may be used to add or subtract these distances. The cost is insignificant in comparison with its value, and the speed of calculations is very greatly increased over that involved in the long method. An experienced slide-rule operator can multiply, divide, square, or get the square root of a number more quickly than the experienced user of printed tables or of "logs" can turn the pages and locate the desired items.

One important limitation of the slide rule is that it will not add or subtract.

A second limitation is that considerable time and practice are required in order to develop speed and accuracy in making calculations. The mastery of slide-rule technique is fully as difficult as that of logarithmic technique. The beginner is likely to misread his rule and make many ridicu-

lous blunders before he has had the necessary amount of practice. All slide-rule work should, therefore, be checked very carefully during the learning stage.

Another limitation of the slide rule is that it yields accuracy only to the third significant digit. It is comparable to logarithms in this respect. This is not at all a serious problem, however, because statistical work seldom requires greater precision than this, and if greater precision is attained it is a case of straining at a gnat after having swallowed a camel, the camel in this case being the error or inaccuracy in the original statistical data. Greater accuracy in slide-rule work can be attained by the simple device of purchasing an extra-length rule. The ten-inch rule will yield accuracy to three places, and the twenty-inch rule, by expanding the scale and introducing finer subdivisions, permits the addition of two more significant digits of accuracy.

In purchasing a slide rule for statistical work it is preferable to get a quite simple one, rather than one with the many elaborate scales which may be of service to engineering students and other technical experts. The authors recommend one which has the A B C D and CI scales. It is well to insist upon the CI scale because of its distinct superiority over the C scale for multiplying. The authors' recommendation would be a Keufel and Esser Company ten-inch rule, number 4053–3, or any other make which has the same arrangement of scales. These may be secured from the leading stationers or dealers in engineering supplies.

It would be out of place in this book to give detailed directions for the use of the slide rule. These may be secured from the printed instruction book which ordinarily accompanies the rule when it is purchased.

Calculating machines. The business world has developed to a high stage of perfection a number of different types of mechanical calculators which are quite useful to the educational statistician. The calculating machine supplements the slide rule or logarithmic table very nicely because it is strong where they are weak. It is very useful for addition, and the more elaborate machines will do subtraction, division, and even square roots. Unfortunately, however, the ordinary machine will perform economically only a small number of the more important calculations which the educational statistician has to make, unless the operator is unusually well trained and possesses more skill in the use of the machine than the average statistics student may reasonably be expected to acquire. To do square root on the calculating machine is laborious indeed as compared with the slide rule or the ordinary table, and some of the less expensive machines will hardly do even such processes as multiplication and division economically.

The student must not make the mistake of assuming that doing addition on a machine guarantees its accuracy. The machine will do its part accurately, beyond doubt, but the danger of striking the wrong key is ever present just as it is in typewriting. Also, it is possible to get a wrong answer by not having the machine clear at the start, and in numerous other ways. In other words, the human or psychologi-

cal element cannot be eliminated, and all work must be checked carefully. Even if the machine is set to print the numbers that are added in, the time required to check this with the original data may be an important factor.

A good calculating machine permits carrying out computations to several decimals, and thus makes for very high accuracy. This accuracy is very much needed in the business world where these machines originated, but is not nearly so important in educational statistics.

Most calculating machines do multiplication much more easily than division. Hence, if it is necessary to divide each of several numbers by the same divisor it is preferable to get the reciprocal of the divisor and then multiply it by each of the numbers, thus avoiding any further division.

A few of the leading makes of calculating machines are the Monroe, Marchant, Burroughs, Dalton, and Comptometer.

(*Note:* It is impossible in this chapter to give further details regarding mechanical calculating machines, but the dealers ordinarily are glad to demonstrate their machines before statistics classes, and the teacher may find it profitable to get in touch with them for this purpose.)

The Hollerith machine. The physical labor of sorting, classifying, and tabulating scores becomes enormous when it is necessary to deal with several hundred or several thousand cases. The mere task of making a frequency table for 4000 scores is one of tremendous proportions. This task is greatly increased if instead of an ordinary frequency table one attempts to make a correlation table for 4000 pairs of scores. The United States Census Bureau has faced this

problem for many years, and has solved it in part by the use of special tabulating machines.

One of the most widely known machines for this purpose is the Hollerith machine. Its work is done by mechanically sorting cards into separate bunches or piles according to the way those cards are punched to represent the data. Thus, a card about six inches long and two and one half inches wide is punched with holes in different positions to indicate the scores or data regarding one school child, or one case of any kind, on each of the tests, measurements, or observations that have been made regarding it. Once the entire set of cards has been punched they are fed into the machine which is set to sort them according to their status as shown in a particular column on the card, and the machine automatically throws them into separate piles with all cards in a given pile alike as regards that particular item. The number in each pile may then be counted in the usual way, after which the cards may again be run through the machine and sorted with respect to data in another column on the cards. Large-scale statistical establishments may also have a second machine to operate in conjunction with the Hollerith machine which does the counting and even records the number, very much in the same way as adding machines record totals. The machines are very expensive, are rented and not sold, and few establishments or institutions can afford their use.

The individual statistician or research worker would seldom be able to afford to employ a Hollerith machine because of its expense and his lack of enough work to keep

it busy. Furthermore, the operation of a Hollerith machine is a somewhat specialized type of skill which the amateur does not have.

The point here is that if a statistical worker knows of the existence of such a machine, and is dealing with a large enough quantity of data to make its use economical, he may arrange with some institution which has one to have the specialist in charge of it run his data through for him, paying him for his time and trouble, and saving his own time and energy for other aspects of the work. For example, a candidate for a doctor's degree had masses of data regarding each of about 1000 school children. He wished to compute several hundred correlation coefficients for all these cases, which would have taken him so long had he done it by the usual method that he would have been unable to finish the job within the time limits available for his work. He planned what he wanted, sent his data to an institutional statistician, and paid for the making of all the correlation tables. In this manner he saved his own time for those aspects of his investigation which did not lend themselves to routine methods of solution.

Printed forms and data sheets. Large quantities of statistical work require the use of specially ruled paper, or particular styles of organization of data upon the page. The casual observer of students of statistics at work is impressed by the fact that they spend quite a bit of time in ruling their paper into rows and columns to take care of the kind of data with which they are working. Probably in an earlier day bookkeepers wasted their time ruling their

paper preparatory to recording their accounts, but eventually they discovered that paper could be more economically purchased from a printer already ruled in special ways for special purposes. Statistical workers are coming gradually to the same conclusion. A student should purchase, at the outset of a statistics course, a supply of paper ruled both ways in rectangles approximately one-half by one-fourth inches. This type of general utility paper will save much valuable time and render a better service than that which the student rules with pencil or ink.

In addition to such general service paper it may be well to purchase, at the appropriate time, a supply of blank forms for such special purposes as computing the coefficient of correlation. A number of such correlation work-sheets are available, such as those of Kelley, Holzinger, Thurstone, Otis, Toops, and others. Also Otis's "Universal Percentile Graph" sheets may be purchased in quantities, from the World Book Company, to shorten the labor of computing percentile ranks and drawing the corresponding curves.

TRUE-FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. If you purchase any instrument for calculation it should be one which you can carry with you at all times.
2. Everything considered, logarithmic tables are the best aids to statistical calculation.
3. The slide rule is more useful than the calculating machine in ordinary statistical work.
4. Tables of squares, square roots, products, etc., have the advantage that they are absolutely accurate.
5. Short methods are likely to be longer than the regular methods at first.
6. Short methods should be checked by long methods at first.
7. Any one with ordinary intelligence can master the use of the slide rule in fifteen minutes.
8. Any student who really has a use for a course in statistics is justified in purchasing a slide rule for his calculations
9. Decimals should never under any circumstances be carried beyond two places in educational statistics.
10. Most tables of products, quotients, etc., require interpolation if numbers are over two or three digits.
11. It is common for students to master the use of logarithmic tables in about an hour.
12. Long methods should be checked by short methods in the later stages of work.
13. Calculating machines are more needed in ordinary statistics than in the business world.
14. The slide rule is better for subtraction than it is for division.

15. Any one who expects to do a statistical type of master's thesis should expect to utilize a Hollerith machine.
16. Labor-saving devices take more time and are really less accurate than the old standard methods of computation.
17. The slide rule enables one to add and subtract large numbers as quickly, but not quite as accurately, as by the usual methods.
18. Logarithms are very useful in multiplication and division.
19. The correctness of computations is not guaranteed by the fact that they are done on a perfect adding or calculating machine.
20. The Hollerith machine is a device for sorting and tabulating certain types of data by means of punched cards.

QUESTIONS AND EXERCISES

1. What is your estimate of the value of short-cut methods in general?
2. What dangers or limitations would you mention that attach to their use?
3. Compare the relative advantages of the four main calculating devices, namely, (*a*) Tables of squares, etc.; (*b*) logarithmic tables; (*c*) slide rule; and (*d*) calculating machines, on the basis of each of the following:
    a. Accuracy
    b. Speed
    c. Expense
    d. Ease of learning how to operate
4. Which of the four types of calculating devices, everything being considered, do you think is most applicable to your own situation, or to the type of statistical work which you expect to do?
5. Discuss the advisability of your college or university's buying a good calculating machine, and making it available to statistics students in some conveniently accessible room.

6. Assuming the price of a good slide rule to be about six or seven dollars, discuss the advisability of making the possession of a slide rule a requirement of all students in a statistics course.

7. Would you consider it wise to purchase a "students' slide rule" for about a dollar, in order to learn how to operate the slide rule, in view of the fact that it is so inaccurate as to be undependable for regular work, making necessary the purchase of a more expensive one if you ever need a slide rule for serious purposes?

8. For practice in using whichever calculating method you are going to use permanently in this course, do the following operations:

$\sqrt{150}$	85/21	42×51
$\sqrt{15}$	8.5/21	42×51.5
$\sqrt{15.5}$	850/21	42.5×51
$\sqrt{155}$	17/73	425×51
$\sqrt{1550}$	17/73.5	422×51
$\sqrt{155.2}$	17.5/73	42×515
$\sqrt{15.3}$	17.2/73	87×31
$\sqrt{15.8}$	17/73.2	87×31.4
$\sqrt{7230}$	49/584	875×31
$\sqrt{9825}$	96/27.3	938×44

9. List the special problems or difficulties met with in doing the above operations, and bring them up for class discussion.

CHAPTER III

TABULATING DATA

WHAT is probably the most fundamental device of statistical practice is known as the frequency distribution. It is simply a systematic and orderly arrangement of the individual facts, measurements, or other data with which the statistician must work. It does for the statistician what the drawer of the cash register does for the cashier by providing definite and appropriate compartments for pennies, nickels, dimes, etc. The frequency distribution is to statistics, in a sense, what the map is to geography, a convenient device to enable one to visualize many facts by grouping them into a unified system. In addition to this general function, it facilitates the computation of certain measurements to be introduced later.

A simple frequency distribution. We may illustrate the nature of the frequency distribution in connection with the following data. A teacher gave a ten-word spelling test to a class of 32 students. The number of words spelled correctly by individual students were as follows: 8, 6, 6, 7, 3, 5, 8, 5, 6, 5, 4, 7, 7, 5, 6, 5, 7, 6, 8, 5, 6, 7, 6, 9, 3, 4, 6, 5, 10, 6, 4, 7. It is difficult to get a very definite idea of the results of this test with the data in this form; hence we tabulate as in Table I.

In tabulating data into a frequency distribution, as in Table I, we first set down in column 1 the numbers which correspond to the scores actually made, with the largest

Table I. Illustration of a Simple Frequency Distribution, with Tabulation
(The data are spelling scores of a class of 32 students)

Scores	Frequency tabulation	Frequencies (f)
(1)	(2)	(3)
10	/	1
9	/	1
8	///	3
7	~~////~~ /	6
6	~~////~~ ////	9
5	~~////~~ //	7
4	///	3
3	//	2
Total		32 (N)

scores at the top, and the remaining scores in order of size. The next step *does not* consist of counting all scores of 10, of 9, of 8, etc., and then filling in the frequency tabulation column 2; on the contrary, the number of errors of tabulation may be considerably reduced if each score is tabulated in the order given. Thus, in tabulating the scores of this spelling test, a tally mark is placed opposite 8; then two such marks are placed opposite 6; one such mark is placed opposite 7, etc. When the tabulation is completed, it will look like column 2, and the total of frequencies (indicated usually by f) in each case may be written opposite the appropriate scores. The total of column 3 (in this case 32) is usually indicated by N, and it designates the total

number of measures or scores tabulated in the frequency table.

Ordinarily, the frequency tabulation (column 2) is not given in a frequency table, but it is included here merely to emphasize the fact that in organizing data into a frequency table each measurement or score should be taken account of in the order given, and tabulated opposite the proper score found in column 1.

Two methods are commonly used to check the accuracy of tabulations: First, the original scores are counted as given, and this number compared with N to see if any scores were missed or counted twice. Agreement here does not guarantee accuracy, however, because some of the scores may have been tabulated in the wrong horizontal line. It is customary to repeat the tabulation using a dot over each tabulation mark (/) and, if the two tabulations agree, the chances are good that the work is correct.

An inspection of Table I reveals a number of facts, among them the following: No student spelled all of the words incorrectly; only 1 student spelled all ten correctly; the lowest score made was 3 words spelled correctly, and this record was made by 2 students. In the same manner, column 3 reveals in each case the number of students who made the scores in column 1. The score of 6 was made by 9 students; 12 students did less well than this, 11 did better, and so on.

Tabulation of more complex data. Most of the data of statistics are not so simple as the illustration given. For example, 55 students took the Terman Group Intelligence

Test, and the following scores represent their intelligence quotients: 114, 107, 114, 114, 109, 111, 106, 97, 112, 139, 107, 109, 97, 111, 132, 99, 118, 107, 96, 106, 123, 115, 116, 120, 115, 86, 112, 109, 115, 97, 104, 100, 128, 126, 96, 115, 110, 111, 110, 111, 95, 102, 106, 100, 112, 83, 114, 110, 85, 110, 91, 83, 85, 128, and 91. An examination of these data shows that the quotients range from 83 to 139, or over a distance of 56 points. A column of 55 entries, with 139 at the top and 83 at the bottom, would be so long and cumbersome that its value would be questionable. Hence a method of grouping scores has been developed and is illustrated in Table II.

TABLE II. ILLUSTRATION OF A FREQUENCY TABLE GROUPED IN CLASS INTERVALS

(The data are intelligence quotients of 55 students)

Scores	Frequencies (f)
(1)	(2)
135–139	1
130–134	1
125–129	3
120–124	2
115–119	6
110–114	15
105–109	9
100–104	4
95– 99	7
90– 94	2
85– 89	3
80– 84	2
Total	55 (N)

The class interval. There are several things about Table II that should receive special attention. First of all, notice the manner in which scores have been grouped. In-

stead of referring to a single score each time, each number in the frequency column now refers to a group of scores. The distances from 80 to but not including 85, usually designated 80–84, and the distance from 85 to but not including 90, usually designated 85–89, etc., are known as "class intervals."

In Table II, there are five scores in each class interval. This number is not the result of chance, but was determined by the application of a definite rule. For convenience of handling and ease of interpretation, it is customary in statistics to group the scores so that the number of class intervals will be not less than 10 (data permitting), nor more than 20. It is obvious, therefore, that the first step in the process is to get the *range* of scores, or the distance between the highest and lowest scores. In the present instance it is from 83 to 139, or 56 points. When this number is divided by 5, the resulting quotient is 11.2, which is more than 10, and less than 20; consequently this was a legitimate number of scores to group in each interval. However, either 3 or 4 would have satisfied this rule, since 56 divided by 3 yields a quotient of 18.66, and 56 divided by 4 gives a quotient of 14.

The rule for finding the size of the class interval may thus be stated as follows: Divide the range by a number such that the resulting quotient will be not less than 10 or more than 20. Any number which satisfies these conditions may be taken as the size of the class interval. Thus if the range of scores is 24, the number 2 is the only one which satisfies the conditions and each class interval will

contain two scores; if the range is 30, either two or three scores may be grouped in each class interval. The rule is not an arbitrary and rigid one, however.

The specification that there should be as many as 10 class intervals is dictated by the need of getting fairly homogeneous groupings, and the idea of keeping the number of intervals below 20 is primarily dictated by the need for economy of labor.

Having found the size of the class interval, the beginning of the first interval should next be determined. When using a range of 5 scores for intervals, each interval should begin on a multiple of 5, as in Table II. In using an interval of 7 scores, each interval should begin on 14, 21, 28, or other multiple of 7, depending on the range of data; and in using class intervals of 4 scores, each should begin on 4, 8, 12, 16, or some other multiple of 4. Any other beginning point would give as accurate mathematical results, but would not be so easy to keep in mind when tabulating.

> *Problem:* At this point, the student should make two additional tabulations of the data of Table II, using class intervals of 3 and 4.

The midpoints of class intervals. The exact nature of a class interval may become clearer if it is illustrated graphically, as in Figure 1.

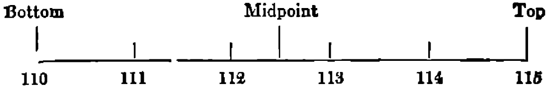


FIG. 1. GRAPHIC ILLUSTRATION OF A CLASS INTERVAL

From the manner in which the class intervals in Table II are indicated, it is easy to be misled in locating the midpoints. There it appears that the midpoint of interval 80–84 is 82; that of interval 85–89 is 87; that of 110–114 is 112. An examination of Figure 1, however, shows that the latter interval begins at 110, and extends the whole distance up to but not including 115. The midpoint of this distance is very obviously 112.5 and not 112.

As soon as each score is tabulated in its proper class interval (see Table II), it loses its identity; we know only that it falls somewhere between the beginning and the end of the interval. In computations to be made later, however, some score must be chosen to represent the interval as a whole, and this score is the midpoint. For this purpose it is assumed that all scores of the class interval are distributed evenly over the entire interval and that the midpoint is the average. When the number of measurements is sufficiently large, this assumption is in general a valid one, but, generally speaking, it becomes less valid as the number of scores becomes smaller; hence it is questionable to group data in a frequency table when the number of cases in the entire group is very small.

This assumption of even distribution of scores over the entire interval has an additional aspect. For example, in the interval illustrated in Figure 1 (scores 110–114 in Table II), it is assumed that there are three scores of 110, 3 of 111, 3 of 112, 3 of 113, and 3 of 114. This assumption is important in certain computations to be made later.

The meaning of a single score. For most data with

which teachers will work, the meaning of a single score is interpreted exactly like that of the class interval. Thus, when single scores are expressed in whole numbers, score 80 of the lowest class interval of Table II is thought of as extending from 80 up to but not including 81, with its midpoint at 80.5; and just as single whole scores lose their identity within the intervals, fractional scores such as 80.1, 80.3, and 80.7 (where such finer measurements are possible) lose their identity within the space between 80 and 81 and are all considered 80's. The explanation for this interpretation is simple: Just as scores of 80, 81, 82, 83, and 84, representing differences in achievement in reading within a class interval, are grouped together for convenience, so may very fine differences in achievement which are theoretically possible between 80 and 81 be grouped for convenience as 80's.

There are other types of data, however, in which the interpretation of scores, particularly the location of midpoints, is contrary to the above plan. Thus, in life-insurance statistics, age 20 does not mean from 20 up to but not including 21; on the contrary, the score named is itself the midpoint of a group of possible ages, such as 19 years, 6 months, 1 day; 19 years, 11 months; 20 years, 5 months, 12 days, etc., extending from 19 years, 6 months, to 20 years, 6 months. This is because insurance companies count ages to the nearest rather than the last birthday. In a similar manner, score 80 on the Ayres Handwriting Scale is usually thought of as any scale value between 75 and 85. (The scores on the Ayres Scale are eight in number: 20, 30, 40, 50, 60, 70, 80, and 90.) Thus the interpretation of a single

score and the location of its midpoint can be made only after an analysis of the manner in which the original data were collected.

Problem: At this point, the student should locate the midpoints of the class intervals of Table II, and also the midpoints of the class intervals of the two additional frequency tables into which these data have been tabulated.

Designation of class intervals. There is a variety of practice in the designation of class intervals, and care is necessary in order to avoid confusion. Thus the same intervals may be expressed in four different ways, as is illustrated in Table III. In practice, all of these methods are employed. While the method illustrated in column 2 is most exact, the method of column 4 probably is the most frequently used.

TABLE III. ILLUSTRATIONS OF METHODS OF DESIGNATING CLASS INTERVALS

(Sample class intervals expressed in four ways)			
Very confusing	Very exact	Less exact, but clear	Commonly used for convenience
(1)	(2)	(3)	(4)
135–140 130–135 125–130 120–125 etc.	135–139.99+ 130–134.99+ 125–129.99+ 120–124.99+ etc.	135–139+ 130–134+ 125–129+ 120–124+ etc.	135–139 130–134 125–129 120–124 etc.

Continuous and discrete data. The discussion concerning the nature and representation of a score or a class interval brings us to a distinction between two main types of data — continuous and discrete. Fortunately, this dis-

tinction is not an important one, especially for beginners in statistics, since most measurements in which teachers are interested are continuous in nature. The distinction, however, exists, and may be mentioned in passing. Data in which there are no fractional units of measurement are discrete; that is, actual breaks occur between units.

For example, the number of persons of each race or occupation in a census table is stated in discrete units. One cannot have $8\frac{1}{2}$ people who are white, or barbers. Again, as our schools are generally organized, a child is in the sixth, seventh, or some other definite grade (though he may be in some division or section of that grade), and not in the 6.72 grade. In the matter of age and in the designation of abilities and traits, however, there are no such breaks theoretically. To be sure, we speak of individuals as being 27, 28, or 40 years old; but in reality the 27-year-old individual is 27 years and some months, weeks, days, hours, minutes, seconds, etc., old. School attendance and grade placement data are designated in discrete units, while age data are essentially continuous.

Frequency rectangles and histograms. In studying data, and particularly in reporting and interpreting them, it is frequently desirable to use a method more graphic or striking than a frequency table. If, for example, we should use rectangles to represent the frequencies of column 2, Table II, the result would appear as in Figure 2.

There are several important facts which must receive consideration in order to grasp the entire significance of this graphic representation of the data of Table II. Notice that

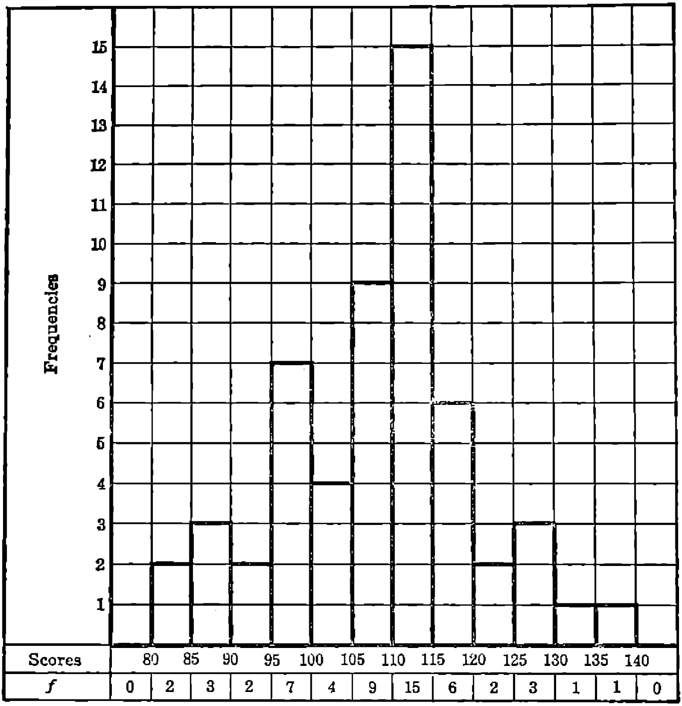


FIG. 2. ILLUSTRATION OF THE GRAPHIC REPRESENTATION OF A FREQUENCY DISTRIBUTION: FREQUENCY RECTANGLES

The data are from Table II.

only the limits of the class intervals have been designated; for example, interval 85–89 means from 85 up to but not including 90. The upper limit of this interval occurs, therefore, where the 90–94 interval begins; hence the beginning of each interval is indicated, but the upper limit of the interval is at the exact point where the next higher interval begins, and does not require separate designation.

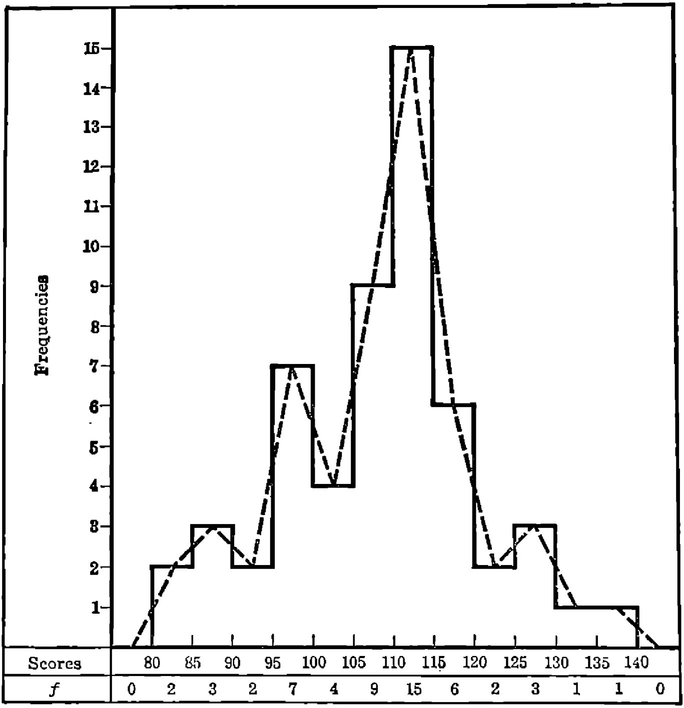
Each rectangle helps us to visualize the frequencies of a particular class interval; for example, in the 80–85 interval, the top of the rectangle is opposite the figure 2 in the column at the left designating frequencies; and in the interval 105–109 the top of the rectangle is opposite figure 9 in the frequency column. Thus it is possible to take in, at a glance, the relative frequencies in the different class intervals of the frequency table.

It is interesting, further, to note that each rectangle begins at the base line and extends to the horizontal line representing the appropriate frequency, and that the sides of these rectangles represent the limits of the class intervals.

By omitting all sides common to the rectangles of Figure 2, we may obtain a figure known as the histogram. The heavy line of Figure 3 constitutes such a histogram. This is a very simple but a very accurate method of graphic presentation.

Frequency polygons and curves. Another method of graphic representation is by means of frequency polygons and curves. In Figure 3, the broken line represents a frequency polygon. It is obtained by taking into account the midpoints rather than the limits of the class intervals, as is the case in the construction of the frequency rectangles or histograms. The midpoint of each class interval is located opposite the appropriate frequency indicated to the left of the figure, and these points connected.

In drawing a frequency curve there is one innovation: Since a polygon is an enclosed area, the frequency-line, the irregular line connecting the various midpoints of the class



Key: ——— Histogram. – – – – Frequency Polygon.

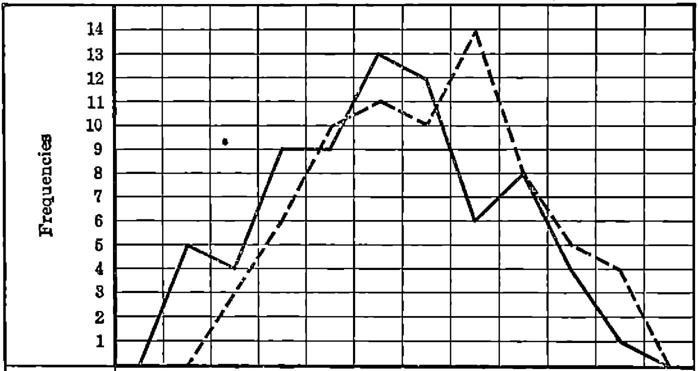
FIG. 3. ILLUSTRATION OF THE GRAPHIC REPRESENTATION OF A FREQUENCY DISTRIBUTION: HISTOGRAM AND FREQUENCY POLYGON

The data are from Table II.

intervals, must be brought to the base line. There being no frequencies in the class interval next below 80–84, and none above the class interval 140–144, it is entirely logical to bring this irregular line to the base line at the midpoint of these intervals. A frequency polygon is always completed in this manner.

When a large number of measurements or scores are tabulated, the frequency line usually is regular and symmetrical. Theoretically, an infinite number of cases with very small class intervals would produce a frequency curve, rather than a frequency line. Irregular frequency lines are often smoothed, and thus represent the probable effect of adding more cases.

Comparison of groups by frequency polygons. The frequency polygon performs an important function in the graphic comparison of two or more groups. For example, a school gave the Haggerty Reading Examination, Sigma 3, to 71 boys near the beginning and again near the end of a semester. The scores were tabulated and the resulting frequency-distributions are presented immediately below



Scores	30	35	40	45	50	55	60	65	70	75	80	
f Test 1	0	5	4	9	9	13	12	6	8	4	1	0
f Test 2	0	0	3	6	10	11	10	14	8	5	4	0

Key: ——— First test, – – – – Second test.

FIG. 4. ILLUSTRATION OF THE COMPARISON OF GROUPS BY MEANS OF FREQUENCY POLYGONS

The data are scores from the Haggerty Reading Examination, Sigma 3, for 71 boys at the beginning and end of the semester.

Figure 4. This figure shows at a glance that the scores of the distribution represented by the broken line (second test) are slightly better than those of the first; that is, there are more high scores and fewer low scores on the second test than on the first, indicating that improvement in reading ability actually took place during the period between the two tests.

Difficulty involved in comparing groups of unequal size by plotting simple frequencies. The Haggerty Reading Examination, Sigma 3, was given to 35 girls at the same time that the 71 boys plotted in Figure 4 were given their second test. The scores of both these tests are presented immedi-

Scores		35	40	45	50	55	60	65	70	75	80
f, Boys	0	3	6	10	11	10	14	8	5	4	0
f, Girls	0	1	1	2	4	5	7	9	5	1	0

Key: ——— Boys, – – – – Girls.

FIG. 5. ILLUSTRATION OF THE DIFFICULTY INVOLVED IN COMPARING GROUPS OF UNEQUAL SIZE BY PLOTTING SIMPLE FREQUENCIES

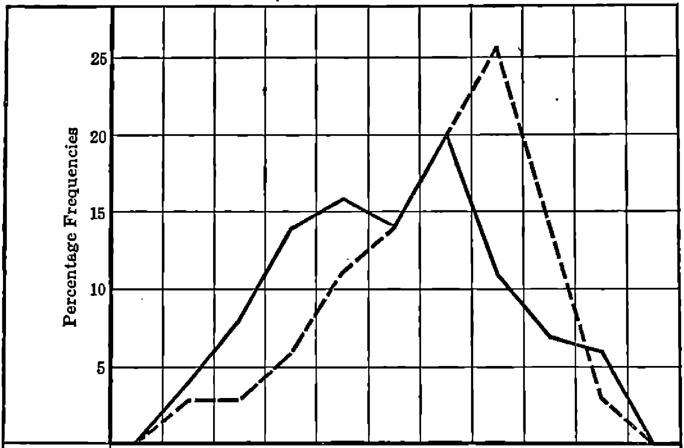
The data are scores of 71 boys and 35 girls from Haggerty Reading Examination, Sigma 3.

ately below Figure 5, and each set of scores has been expressed as a frequency polygon. These polygons, unfortunately, are not of equal size (do not enclose equal areas), and the resulting comparison is not at all satisfactory. In fact, as will be evident later, plotting polygons directly from frequencies, when the numbers of cases differ, results in a definite deception; and the greater the difference in total cases, the greater is the deception. In Figure 4, the number of cases represented in each polygon was equal; under those conditions, frequency polygons may always be plotted directly from the simple frequencies, but the greater the difference in number of cases, the greater is the need for another method of comparison.

Advantage of comparing groups of unequal size by plotting percentage frequencies. The difficulty experienced in comparing groups of unequal size, when simple frequencies are plotted as in Figure 5, may be obviated by reducing each simple frequency to its per cent of the total value, and then plotting the resulting percentage frequencies. In Figure 6, the data of Figure 5 have been repeated, but immediately before the score-frequencies for both boys and girls there has been indicated the percentage frequencies, which in each case are obtained by dividing each frequency by N.

Of the 71 cases representing the total number of boys, 4 occur in the interval 75–79. In order to determine what percentage of the total 71 cases occur in this interval, we must divide 4 by 71, and the resulting quotient is 5.63 per cent (roughly 6), which is placed below the number represent-

ing the simple frequency. In the distribution of girls' scores, one case occurs in the interval 75–79. To determine what per cent this frequency is of the total of 35 frequencies in the column, we divide by 35, and the resulting quotient of 2.86 per cent (roughly 3) is placed below the number representing the simple frequency.



Scores	35	40	45	50	55	60	65	70	75	80	
f, Boys	0	3	6	10	11	10	14	8	5	4	0
% *f*	0	4	8	14	16	14	20	11	7	6	0
f, Girls	0	1	1	2	4	5	7	9	5	1	0
% *f*	0	3	3	6	11	14	20	26	14	3	0

Key: ——— Boys, – – – – Girls.

FIG. 6. ILLUSTRATION OF THE COMPARISON OF UNEQUAL GROUPS BY PLOTTING PERCENTAGE FREQUENCIES

The original data are the same as in Figure 5.

Considerable time may be saved if, instead of dividing each simple frequency by *N* (the total number of cases) in order to find the corresponding percentage frequency, the reciprocal of *N* is first found and this value multiplied by

each score-frequency. In the present case the reciprocal of 71 is .01408 +. Multiplying this by 4 (the number of frequencies in the 75–79 interval), we obtain a product of 5.63 + per cent. Multiplying the girls' frequency 1 by the reciprocal of its N (.02857 +) yields a product of 2.86 per cent which is its corresponding percentage frequency. Ordinarily one would not retain the decimals of percentage calculation, since percentages themselves represent two decimals, but the decimals for the reciprocal of N must be preserved until the multiplication has been completed.

In Figure 6, a scale of percentage frequencies has been substituted for the simple frequency scales of previous figures, but the location of percentage frequencies for purposes of plotting them is made exactly as before. An examination of the two frequency polygons of Figure 6 reveals at once a superiority in the achievement of girls which was not at all evident in Figure 5, where the polygons were plotted from simple frequencies.

TRUE-FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. A "class interval" means a portion of the range in which no cases happen to fall.
2. Converting frequencies into percentages is simply a means of drawing two graphs to the same scale.
3. Data showing numbers enrolled in each school grade would be classed as continuous rather than discrete.
4. The use of frequency tables in treating data tends to cultivate habits of carelessness and inaccuracy.

5. Ordinarily class intervals should not be less than ten nor more than twenty in number.
6. Frequency polygons are plotted by locating dots for midpoints of intervals.
7. The largest score should always appear at the top of a frequency table.
8. It is assumed that all measures in a class interval are concentrated at the midpoint.
9. Groups of about equal size may be compared by plotting simple frequencies.
10. Most educational data are continuous.
11. If the count of the original scores agrees with N, the tabulation is correct.
12. Intelligence scores should always be grouped in class intervals of five.
13. The frequency table would be useless if you did not plan to base later calculations upon it.
14. Frequency polygons are preferable to histograms if two groups of data are to be shown on the same graph.
15. It would be possible for three different people to get three different frequency tables from the same data and yet all be right.
16. The true meaning of a single score is usually .5 higher than the number which we use ordinarily to designate it.
17. Histograms are the same as frequency rectangles.
18. A frequency table gives mathematically wrong results if the scores are arranged with the smallest at the top and the largest at the bottom.
19. A frequency table with intervals of more than one is almost certain to misrepresent the data slightly.
20. A score of 80 always means from 80 up to but not including 81.
21. Scores should be tabulated in a frequency table in the order in which they appear in the original data.
22. For a class interval designated 85 to 89 inclusive the correct midpoint would always be 87.5.

23. Simple frequencies may be reduced to per cents by multiplying each by the reciprocal of N.
24. The greater the inequality in size of groups to be compared, the greater is the need for using percentage frequencies.
25. A frequency polygon is a graphic representation of a frequency table.
26. Percentage frequencies, when plotted, give a different shape of graph from simple frequencies.

QUESTIONS AND EXERCISES

1. What is a frequency table? A histogram? A frequency polygon? A frequency curve?
2. What are the important steps in expressing data in the form of a frequency table?
3. What is meant by range? What is a class interval?
4. In what different ways may class intervals be indicated? Which do you prefer? Why?
5. In tabulating data into a frequency table, how is the number of class intervals determined?
6. In plotting polygons from frequency tables for purposes of comparison, when may one plot the score frequencies, and when should one not plot percentage frequencies?
7. The following data represent the total days' attendance of each of 60 pupils during a school year: 163, 145, 125, 162, 174, 165, 141, 123, 132, 173, 173, 174, 176, 144, 156, 157, 156, 162, 161, 161, 161, 148, 173, 159, 126, 148, 142, 137, 160, 168, 163, 155, 155, 141, 179, 153, 179, 152, 136, 167, 151, 177, 164, 156, 164, 154, 171, 163, 164, 163, 154, 175, 167, 161, 156, 162, 166, 172, 153, 132.
   a. Tabulate these attendance records into a frequency table.
   b. Express the frequency table in the form of frequency rectangles. Convert it into a histogram by means of heavy lines.
   c. On the figure already completed, superimpose a frequency polygon.

8. In a certain Los Angeles high school the following scores were made on the Brown-Woody Civics test by the 12B June class of 1927: 112, 121, 92, 103, 88, 97, 116, 138, 107, 113, 133, 126, 126, 96, 108, 112, 113, 122, 123, 103, 118, 117, 123, 111, and 113. In June 1928, the 12B Civics class in the same school took the same test, making the following scores: 83, 138, 124, 116, 118, 113, 106, 103, 108, 96, 93, 85, 111, 123, 131, 132, 115, 124, 114, 133, 115, 113, 134, 118, 121, 106, 133, 108, 114, 107, 124, 116, 106, 128, 118, 121, 116, 126, 103, 119, 102, 126, 118, 100, 122, 138, 115, 113.
   a. Express each group of scores in the form of a frequency table using the same class intervals.
   b. Express each table in the form of a frequency polygon, plotting the score frequencies.
   c. Express the frequencies of each table as percentages.
   d. Plot the percentages of the 1927 class as a frequency polygon.
   e. Superimpose upon the frequency polygon representing the 1927 scores a similar polygon representing the 1928 scores.

9. Tabulate the following measures into three frequency tables, using 2, 3, and 4 respectively, as class intervals:

   70, 40, 50, 35, 52, 41, 56, 38, 45, 43, 59, 37, 54, 44, 56, 41, 47, 55, 66, 38, 54, 50, 60, 43, 52, 38, 66, 44, 57, 49, 46, 55, 50, 54, 63, 47, 52, 40, 54, 72, 43, 49, 52, 41, 54, 54, 69, 44, 74, 56, 51, 63, 54, 68, 45, 60, 55, 62, 65, 72, 45, 73, 54, 60, 47, 63, 55, 49, 71, 54, 51, 62, 52, 57.

CHAPTER IV

THE MEAN

THE term "mean" is used advisedly as the title for this chapter instead of the term "average," because there are several different kinds of averages, of which the mean is only one. It is true that the mean is the most commonly used average, and the only one with which many people are familiar. To the statistician, however, the term "average" is quite indefinite and unsatisfactory, and should be replaced by a definite statement as to *which* average is meant, lest there be uncertainty as to whether it refers to the mode, or median, or ordinary mean, or even the harmonic mean, the geometric mean, or various special types of weighted averages.

Need for some kind of measure of central tendency. The mean, as well as the other kinds of averages, has been devised to describe the general trend, or central tendency, of a number of cases. Thus a group of children may be of different ages, but when we say the mean (or other average) age is ten we know that ten is quite typical or representative of the group as a whole, even though there may be some children six years old and others sixteen.

That there is daily, if not hourly, need for some way of representing the central tendency of cases may be illustrated by citing a few expressions from ordinary speech which convey this meaning. For example, we have: "As a general rule," "On the whole," "By and large," "Ordinarily,"

"Usually," "On the average," "In most cases," "The majority," "Most people," "The average child," etc. These expressions indicate that we are frequently making some sort of general statements about particular cases. The procedures of the statistician in computing the mean are simply more exact and precise mathematical methods of doing this than those used by the ordinary person who has no special training in statistics; they are much more precise than many of the guesses, conclusions, or generalizations which the untrained person is likely to express in sentences which begin with the words "On the average..."

Definition of the mean. Probably the most helpful definition we can give for the mean, at this point, is that it is the ordinary "average" with which we are all familiar. It is the quantity found by adding up all the figures or scores and dividing the total by their number. Thus if there are 56 scores to be averaged and their total is 943, their mean is 943/56, or approximately 17.

A somewhat more exact and specific definition of this universally familiar quantity is that *the mean is that point in the range above and below which there is an exactly equal number of units of deviation.* Thus, if one student's score deviates ten points above the mean it must be balanced by one which deviates ten points below it, or by ten scores that deviate one point each below it, or by some other combination, such as three scores that deviate three points below and one that deviates one, etc.

Limitations of the usual method of computing the mean. The expert statistician seldom finds the mean by adding up

grows out of the common tendency to overlook the *midpoint concept* when using ungrouped data. Thus in averaging ages of school children the ordinary person without statistical training will treat age 15 as exactly 15, while a statistician will usually treat it as 15.5, since ages are ordinarily reported on the basis of the last birthday. In this particular the frequency table mean is more likely to be right than the mean derived from ungrouped data, because there is greater probability of making the correct decision as to what is the true midpoint of a score.

TRUE–FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. The mean is an imperfect representation of the data from which it is derived.
2. The terms "mean" and "average" are used synonymously in this text.
3. The same formula should be used in finding the mean whether the class interval is one or larger.
4. In ordinary educational data you must use intervals of one, or you cannot be sure the mean is absolutely correct.
5. Using intervals of one protects the mean against errors introduced by chance factors of grouping.
6. The usual method of finding the mean often yields a result .5 too small.
7. If you get the same mean by two different methods it is certain to be correct.
8. The short method of computing a mean is slightly more accurate than the long method.
9. The terms "mean" and "average" indicate that a generalization has been made with respect to data.

10. No matter what assumed mean is chosen the same result will be obtained if the computation is correct.
11. The short method of computing the mean can only be used with data grouped in intervals larger than one.
12. The mean is a device for economy in thinking about data.
13. The assumed mean should be chosen in the midpoint of a class interval.
14. The mean from a frequency table will ordinarily be slightly different from that derived from ungrouped data.
15. The mean of a group of students is the person who falls at the middle of the distribution.
16. It is impossible to be sure of the mean without knowing something about the nature of the original data.
17. Using less than ten or more than twenty class intervals will detract from the accuracy of the work.
18. The mean is that point in the range of a distribution above and below which there are exactly equal numbers of cases.
19. Some of the work involved in computing the mean by the short method facilitates the computation of other measures.
20. The mean may be found by the short method by starting with an assumed mean.

QUESTIONS AND EXERCISES

The following are the scores of individual students on one section of an intelligence test: 40, 54, 40, 34, 52, 106, 73, 59, 69, 66, 100, 79, 89, 86, 62, 36, 41, 40, 41, 40, 60, 82, 73, 41, 49, 54, 46, 108, 66, 72, 61, 53, 58, 62, 42, 54, 99, 52, 62, 64, 81, 99, 43, 61, 90, 63, 89, 78, 82, 35, 64, 62, 54, 72, 68, 88, 52, 49, 67, 79, 48, 43, 71, 43, 39, 62, 42, 69, 101, 84, 58, 69, 83, 91, 48, 57, 75, 89, 91, 63, 102, 71.

1. Add them and find the mean by the usual method, being sure to check the accuracy of the work very thoroughly.
2. Make a frequency table for the data using intervals of 5, and check the tabulation by retabulating, putting a dot above each tally mark as you tabulate the second time.

3. Find the mean based on the frequency table just finished. Compare it with that derived from the ungrouped data.
4. Check the work in No. 3 above by using a different assumed mean one interval higher.
5. Check the work in No. 3 above by using an assumed mean two intervals lower.
6. Tabulate the original data again using a class interval of 8, and compute the mean. Compare the size of the mean this time with those found by the simple arithmetical method, and by the method using class intervals.
7. Check the work in No. 6 above by using a different assumed mean.
8. Compute the mean for each of the sets of data in the exercises 7, 8, and 9, at the end of Chapter III.

CHAPTER V

THE MEDIAN

Definition of the median, and comparison with the mean. The median resembles the mean in that it is a form of average, or measure of central tendency. It may be used to describe or stand for all the cases in the distribution, of which it is a reasonably typical representative. It will be higher than some of the cases and not so high as others, but is still the best or close to the best single measure which is available to sum up all the cases into a single figure. So far, it resembles the mean, but beyond this point differences appear.

The median differs from the mean as regards the way it is selected to represent the individual scores in the distribution. Just as the popular vote for the President of the United States may differ radically from the vote in the electoral college, due to the difference in the basis of representation, so the median may be quite different from the mean because it is computed in a different manner.

The mean takes account of the *amount* of deviation of each case from the central tendency, whereas the median considers only the *number* of cases that deviate from it. Thus an extremely high or low score affects the mean markedly, while its extreme deviation would not influence the median at all.

For those who wish a concise statement of the difference

between the two measures, the following definitions will be useful:

1. The median is that point in the scale above and below which are found an equal number of cases.

2. The mean is that point in the scale above and below which are found an equal number of units of deviation.

Let us illustrate this difference in terms of the ages of a small group of high-school students. Suppose they are 14, 15, 16, 16, 17, 18, 18, 19, and 29. If we use the mean as our average, the one adult student has considerable weight, and the result is a mean age of 18 years. If, however, we take the median as our average, the adult counts no more than if his age had been 18 or 19 instead of 29, and the median is 17 years, because by inspection of the scores we see that there are four students older than 17 and four younger.

Finding the median of data grouped into a frequency table. The student is likely to have less confusion regarding the true nature of the median if he learns first to compute it from grouped rather than ungrouped data. Let us, therefore, turn our attention to Table VII and see how the median is found.

Since the median has been defined as that point in the scale above and below which there are exactly equal numbers of cases, the first step is to find what number constitutes half the cases. Since N is 94 in this table, there should be one half of 94, or 47 cases, on each side of the median. If, therefore, we start at the bottom and add up the fre-

TABLE VII. ILLUSTRATION OF THE METHOD OF FINDING THE MEDIAN OF DATA GROUPED INTO A FREQUENCY TABLE

(The data represent the number of arithmetic problems solved in a time-limit test.)

Problems solved	f
(1)	(2)
45–49	1
40–44	2
35–39	5
30–34	7
25–29	11
20–24	18
15–19	21
10–14	17
5– 9	10
0– 4	2
Total	94

$$\text{Median} = 15 + 18/21 \times 5 = 15 + 4.29 = 19.29$$

quencies, we find that the forty-seventh case falls in the interval 15–19. We know, then, that the median is somewhere between 15, the bottom of the interval, and 20, the top of it. But just where is it, exactly? We shall see.

If we count the cases up to the bottom of the 15–19 interval we have a total of 29 cases, showing that we must go farther before we reach the median. If we go on to the top by adding in the 21 cases found in this interval, we have a total of 50 cases, showing that we have gone a bit too far. To be exact, we have gone three cases too far, or three twenty-firsts of the interval too far. Since we needed only 18 more cases to make the required 47, we should go only 18 to find the exact median. Since the interval covers a range of 5 scores and we wish to go eighteen twenty-firsts of that distance, we find 18/21 of 5, which is 4.29. We then add this

amount to the bottom of the interval, or 15, and get 19.29 as the true median, or the point which divides the 94 cases into two groups of 47 each.

Steps in the process. Expressing what we have done as a set of detailed directions, we have the following as the steps in the process:

1. Divide N by two (yielding 47 in Table VII).

2. Add up the cases from the bottom until the interval is found which contains the median (15–19 in Table VII).

3. Find how many cases we are required to go in this interval to equal $N/2$ (18 cases in Table VII).

4. Note the total number of cases in this interval (21 in Table VII).

5. Divide the number of "cases to go" by the number of cases in the interval (18/21, or .857 in Table VII).

6. Multiply this quotient by the size of the class interval ($5 \times .857$, or 4.29, in Table VII).

7. Add this product to the score which falls at the bottom of the interval ($15 + 4.29$, or 19.29 in the case of Table VII).

Putting what we have described into semi-mathematical form we have the equation shown under Table VII, which may be used as a model by which to find the medians of other distributions. We may, however, generalize it a bit, and arrive at the following formula:

$$\text{Median} = \text{Bottom of interval} + \frac{\text{cases to go}}{\text{cases in interval}} \times \text{size of interval.}$$

Problem 1: For practice the student should add ten new cases to the frequency for the interval 20–24, keeping the en-

tire table the same otherwise, and compute the median anew for the 104 cases. This will throw the median over into the next interval, and give practice in the same processes as before but with new figures.

Problem 2: For further practice in dealing with an odd number of cases, add one new case in interval 10–14, thus making 95 cases in all, and compute the median as before. The only difference is that the fraction .5 must be used in certain cases. For example, $N/2 = 47.5$ instead of 47, and so on.

Finding the median of ungrouped data. Where there are only a few cases in the group, the making of a frequency table is not necessary, and often not even desirable. To make a frequency table for ten cases may distort the facts in the case, because there may be only one case in a given class interval and it may chance to fall at the very bottom, while the assumption which goes with grouped data is that all cases are distributed evenly throughout the interval. Before data are grouped into a frequency table there should be enough cases to permit a reasonable operation of the law of probability.

For this reason there is need for a method of finding the median when there is only a small number of cases. This is very simple, and involves the following steps:

1. Rank the data according to size.

2. Choose the score which has the midmost rank, which is the median score.

3. In case there is an even number of cases in the group, the median will fall half-way between the two middle scores.

Let us illustrate with two sets of data, one having an odd

number of cases and one having an even number. We shall use very small numbers of cases for convenience.

Class I

Scores	5	8	13	22	29	
Ranks	1	2	3	4	5	

Class II

Scores	3	7	13	15	21	30
Ranks	1	2	3	4	5	6

In Class I the median is the score which has rank 3, namely 13. If there had been 17 cases it would be the score which had rank 9, or if 21 cases, rank 11.

In Class II, the median is half-way between the scores which have ranks 3 and 4, or at score 14. Similarly, if there were 12 cases in the group the median would be the point half-way between the scores which have ranks 6 and 7. This half-way point between the two middle scores may be a whole number or a fraction, depending on whether there is an even or an odd number of units between them.

Various complications enter into the problem of finding the median from ungrouped data, which will not be discussed here. There are many special cases which are a bit puzzling, and it is not uncommon for a class of 20 students to get four or five different medians from the same set of data, with a fair degree of plausibility for each result. The point is that if extreme accuracy and precision are to be insisted upon there should be enough cases to warrant refined methods, and in that case the frequency table should be used.

When to use the median, and when the mean. No hard and fast rule can be laid down to govern the use of the mean and the median. Which one to use must be decided on the basis of a study of the particular situation and an analysis of the factors involved in it.

If the data are "normal" — that is, if they are evenly divided and make the familiar bell-shaped curve when converted into a frequency distribution — the mean and the median will be the same, and the choice between them will depend on the ease of calculation or upon the method that has been employed with other data in the particular investigation. Incessant switching back and forth from one to the other in the same study, however, without apparent reason, is to be avoided.

When the distribution is decidedly skewed — that is, bunched or elongated toward one end or other of the distribution — the mean will differ from the median because the few cases which fall a long way out toward the end of the range will have greater influence on the mean than on the median. In some instances they should have more weight, and in others they should not, depending somewhat upon the philosophy which rules one's thinking on the particular problem.

The following are some instances in which it would seem that the median should be used because the extreme cases ought *not* to be allowed to have extra weight:

1. The median should be used in averaging the ages of college students. The reason is that a few adults of 45 or 50 years of age attend college, and one such person may

counteract the influence of many precocious boys and girls who enter a few semesters earlier than normal if the mean is used.

2. In averaging the heights of a small number of children preparatory to determining the height for a drinking fountain one extremely tall over-aged pupil may raise the mean enough to cause the fountain to be put an inch or two too high for the rest of the group.

3. In deciding public questions which depend upon size of income, the income of a millionaire should not receive proportionate weight, else we should have the interests of millions of citizens outweighed by those of a handful of rich men.

One point which should be called to the attention of the beginner in statistics is that the choice between the median and the mean should be made on the basis of the real merits of the two, and that the plan of finding each and using the one that gives the highest or most favorable result deserves the most vigorous condemnation.

The mode. There remains to discuss a third type of average which is often used in educational literature, but which is not nearly so useful as the mean or the median. It is the mode, or the interval in the range at which there is the largest number of cases. Thus if there are more boys of age 15 in a class than of any other single age, 15 is the modal age for that class. Sometimes a case occurs in which there are two peaks in the distribution curve, or two scores which have an outstanding number of cases, separated by

intervals which have fewer cases. Such a distribution is spoken of as "bimodal." The method of computing the approximate mode is very simple, and consists of making a frequency distribution and picking out the class interval which has the highest frequency. The midpoint of this class interval would then be considered as the mode. For special purposes it might be desirable to compute a more exact mode, but we shall not consider that question here.

The drinking fountain example, given above under the discussion of the median, is an example of a type where the use of the mode might give even better results than the use of the median.

TRUE-FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. The median is always a point and not a score.
2. The median derived from ungrouped data is more accurate than that from grouped data.
3. The median must always be computed from the bottom of the distribution.
4. The median derived from ungrouped data is likely to be too small by exactly .5.
5. In comparing groups as to their averages, that average should be used which shows the largest difference.
6. The median of ungrouped data is less likely to be correct than is the mean of the same data.
7. A bimodal distribution is one which has two medians.
8. The median of ungrouped data is always found between the two middle cases.

9. One should never use the median if the mean can be made to serve the purpose at all.
10. The median is one kind of mean.
11. The median takes account of the number but not of the relative sizes of scores.
12. The mode is of little value because it is almost always the same as the median.
13. The median and mean are the same if the distribution is symmetrical.
14. The mean is ordinarily slightly higher than the median.
15. It is of doubtful wisdom to try to compute the exact median from ungrouped data.
16. No mathematical formula is available for the computation of the median.
17. The median is a less accurate statistical quantity than the mean.
18. The median is always the middle case or score.
19. The short method of finding the median is slightly less accurate than the long method.
20. If thirty classes were arranged according to mean intelligence, and rearranged according to median intelligence, the ranks would be the same throughout.

QUESTIONS AND EXERCISES

1. Find the median of each of the sets of data in exercises 7, 8, and 9 at the end of Chapter III.
2. Compare these with the means of the same data, and explain the reasons for any differences between the two in each instance.
3. In each instance tell which one you think you would use to represent the data, and why.
4. Find the median for the intelligence scores given in the exercises at the end of Chapter IV, with the data tabulated in intervals of 5.

5. Find the median for the same data when tabulated in intervals of 8. Explain the difference in size of median this time, as compared with the previous time.

6. Which would be the best measure to use — mean, median, or mode — in each of the following cases:
   a. Intelligence scores of a class of 40 pupils, in a school having a normal distribution of ability.
   b. Determining the composite price per pound of 40 foods, ranging from potatoes and flour to butter and beefsteak.
   c. Determining the height above the floor to build the top of a make of kitchen ranges.

CHAPTER VI

PERCENTILES

Meaning of percentiles. The small boy who boasted that in his spelling class he had been either head or next to head for three months, when in reality there had been only two students in the class during that time, illustrates the need for a standard method of dealing with ranks. To be next to head in a class of two is quite different from such a position in a class of thirty or forty students.

A great deal of difficulty is experienced when one tries to describe the status of a given pupil in his group, even if the size of the group is stated. For example, Mary has a rank of 17 in a group of 28 students, while Helen has a rank of 27 in a group of 41 students, and their parents wish to know which has the higher rank. The answer is not readily apparent from inspection, and the question as to which is superior may go unanswered. The answer is quite clear, however, if each girl is located according to the rank she would have if her group contained 100 students. Thus Mary's rank would be approximately 59, and Helen's would be approximately 65. To compare ranks 17 and 27 may be misleading, but their percentile ranks are clear and unmistakable in meaning.

We may say, therefore, that the fundamental service of percentiles is to facilitate the study of particular persons or cases in a group by reducing the rank of each in his group to the basis of what it would be if the group contained

exactly 100 cases. Once the percentile rank of a given case has been determined, the number of cases in the group can be dropped from attention, since it serves no useful purpose. This dropping of excess baggage makes for speed, economy, and clearness in thinking about the significance of the data under consideration.

High ranks signify high scores. Let us note here that in statistical work it is advisable to let a high rank be represented by a high number, and not by a low one, as in popular practice. Thus the highest person in a group of 58 students has rank 58, and not rank 1. Consistent following of this standard practice will save much confusion and error.

Relation of percentiles to median, quartiles, quintiles, and deciles. In Chapter V we studied the median. We noted that it was that point in the range above and below which there fell exactly the same number of cases. In other words, the median is the 50-percentile, since it is the point which divides the lower 50 per cent of all the cases from the upper 50 per cent. Likewise, the lower quartile is the 25-percentile, and the upper quartile is the 75-percentile.

In a similar fashion we may speak of quintiles (the points which divide the total number of cases into fifths), or deciles (the points which divide them all into tenths). The percentiles simply go further and divide them into hundredths.

Definition of percentiles and of percentile ranks. If the entire set of scores in a distribution is thought of as arranged side by side in order of size, and the entire row or series divided into one hundred equal parts, the points which separate these little hundredths may be properly called per-

centiles. A percentile is therefore a *point* and not a *distance* on the line, when the cases are arranged in order of size.

When we speak of the *percentile rank* of a given score or case we mean simply the percentile nearest to which that score or case falls in the distribution. Thus, if a score of 23 is nearer to the 65th percentile than to the 64th we say it has a percentile rank of 65.

Educational situations illustrating the use of percentiles. In order further to clarify the meaning of percentiles, let us illustrate with a few concrete cases, as follows:

1. In a university 947 freshmen entered in the autumn and took the required intelligence test. In the spring 321 new freshmen entered and took a different test. The test scores themselves are not comparable, nor are the simple ranks, but if each student's position in his group is translated into that rank which he would have if exactly 100 students had taken the examinations, a percentile rank of 88 means the same position in the series in the autumn as in the spring, signifying in each case the point in the total range of scores which separates the upper 12 per cent from the lower 88 per cent of the group. It is then possible to make calculations including members of both groups in a single distribution because the percentile ranks constitute a sort of "common denominator."

2. A parent wishes to know whether to invest money in the musical education of his daughter. He has her tested for musical talent by means of the Seashore test. He is told that she has a score on the pitch-discrimination test sur-

passing 93 per cent of the school children who have taken the test (percentile rank 93), that she surpasses 96 per cent (has percentile rank 96) in loudness discrimination, and has similarly high scores for other parts of the test. The superior talent of his daughter is thus very concretely grasped by the parent and he concludes that it is worth while to invest in music lessons for her. Furthermore, the specialist who gave the test can determine this child's percentile rank simply by referring to the table of scores from persons previously tested, reading off immediately the percentile rank that corresponds to a score of the size which she earned.

The cumulative-frequency table as an aid to computation. In most statistical studies the number of cases is so large that the use of a frequency table is a source of considerable economy in computing percentiles. A new principle enters at this point, however, which has not been considered before in this book, namely that of *cumulative frequency*.

Let us explain what is meant by reference to Table VIII. There is nothing new in the first two columns, which are like any ordinary frequency table. The third column, however, headed "*cum. f.*" for "cumulative frequency," is arrived at by adding each frequency to the total of all frequencies below it. Thus there is one case in the lowest interval, for scores 0–2, and there are three cases in the next one above it. The cumulative frequency for the lowest interval is the same as the regular frequency, namely, 1; but that of the second interval is 3 higher, since the simple frequency for the second interval is added to that of the one

TABLE VIII. ILLUSTRATION OF THE CUMULATIVE FREQUENCY TABLE

(The data are completion-test scores of 161 college freshmen.)

Completion test scores	*f*	*Cum. f*
(1)	(2)	(3)
30–32	1	161
27–29	3	160
24–26	7	157
21–23	14	150
18–20	23	136
15–17	35	113
12–14	31	78
9–11	26	47
6– 8	17	21
3– 5	3	4
0– 2	1	1
Total	161	

below it. The numbers in the cumulative-frequency column thus get larger and larger as they approach the top of the table, since each interval involves adding in an additional number of cases. The final interval at the top has a cumulative frequency of 161, or exactly the same as N, since all cases in the group have been counted by that time. In other words, the cumulative-frequency column is constructed simply by adding up the ordinary frequencies, writing down the total each time a new interval is added in.

Problem: The student should make a cumulative-frequency table at this point for practice, using data from regular frequency tables found elsewhere in this book. This practice should precede the study of the next step which introduces the translation of cumulative frequencies into percentiles.

Converting cumulative frequencies into percentile ranks of the tops of class intervals. Just as simple frequencies

may be expressed in terms of per cents of the total number of cases (see Chapter III), cumulative frequencies may also be turned into percentage form by dividing each cumulative frequency by *N*. Thus, referring to Table IX (which is similar to Table VIII except for having new data), starting at the bottom interval we have a cumulative frequency of 2. Dividing this by *N*, or 131, we have approximately 2 per cent (1.525 to be exact), indicating that when we reach the top boundary of this class interval we have passed 2 per cent of the cases. The division of 5 by 131 gives us 4 per cent, or the percentage of cases passed when we reach the top of the second class interval. Similarly, the third interval, with its cumulative frequency of 14, would become 14/131, or 11 per cent, and so on for the remaining intervals.

TABLE IX. ILLUSTRATION OF THE PROCESS OF CONVERTING CUMULATIVE FREQUENCIES INTO PERCENTILE RANKS OF THE TOPS OF CLASS INTERVALS

(The data are scores of 131 college freshmen on an intelligence test.)

Intelligence test scores	*f*	*Cum. f*	Percentile ranks of tops of intervals
(1)	(2)	(3)	(4)
110–119	1	131	100
100–109	2	130	99
90– 99	2	128	98
80– 89	7	126	96
70– 79	20	119	91
60– 69	21	99	76
50– 59	38	78	59
40– 49	26	40	30
30– 39	9	14	11
20– 29	3	5	4
10– 19	2	2	2
Total	131		

The work is facilitated if the reciprocal of N is found and multiplied by the cumulative frequencies.

Problem: The student should pause here to convert the cumulative frequencies in Table VIII into per cents, in the same way as those in Table IX.

Finding percentile ranks of the midpoints of class intervals. The reader should note that the cumulative frequencies in Tables VIII and IX represent the *tops* of the class intervals, and not the midpoints. Thus in Table IX the percentile point 76 corresponds with the score 69.999 rather than the midpoint, 65. Similarly, the percentile 100 matches the score 119.999 rather than 115, the midpoint. The result is that whereas the percentile column runs up to 100 per cent at the top, it only goes down to 2 per cent rather than 0 per cent at the bottom. Clearly, therefore, it is necessary to devise some way to compute the percentile ranks of the midpoints, else we shall always have percentile ranks one half an interval larger than they actually are.

There are several ways of finding the percentile ranks of the midpoints after the percentile ranks of the tops — and therefore of the bottoms — have been found. One is to draw the percentile graph (as shown later in this chapter) and find the percentile rank of each midpoint from the graph. This will become clear later.

A fairly simple way to get the percentile ranks for the midpoints mathematically is by interpolation, or finding the number halfway between the percentile for the bottom and that for the top of the interval. This may be done by

inspection, without great difficulty, but the accuracy of the work is improved if another column is added to the table to show the increase in the percentile rank of the top of each interval over that of the bottom. After this increase has been determined, it is a simple matter to subtract one half that amount from the percentile rank of the top of the interval, which gives the percentile rank of the midpoint.

The procedure illustrated. Let us illustrate this by reference to Table X. The first four columns are the same as in previous tables except that they represent different data.

TABLE X. ILLUSTRATION OF THE METHOD OF FINDING THE PERCENTILE RANKS OF THE MIDPOINTS OF CLASS INTERVALS
(The data are scores of 486 college freshmen on a grammar test.)

Grammar test scores	*f*	*Cum. f*	Percentile ranks of tops	Increase	Percentile ranks of midpoints
(1)	(2)	(3)	(4)	(5)	(6)
85–89	5	486	100	1	99.5
80–84	9	481	99	2	98
75–79	23	472	97	5	94.5
70–74	52	449	92	10	87
65–69	81	397	82	17	73.5
60–64	103	316	65	21	54.5
55–59	74	213	44	15	36.5
50–54	48	139	29	10	24
45–49	37	91	19	8	15
40–44	22	54	11	4	9
35–39	16	32	7	4	5
30–34	7	16	3	1	2.5
25–29	5	9	2	1	1.5
20–24	4	4	1	1	.5
Total	486				

The fifth column shows the increase in percentile rank for each interval, arrived at by subtracting from each entry the amount of the one below it. Thus the percentile rank

of the top of the interval 20–24, or of the score 24.999, is 1, an increase of 1 over the interval below it, which was zero. Similarly, the score 29.999 shows an increase of 1 over 24.999, and the score 74.999 shows an increase of 10 over 69.999.

When column 5 is complete, column 6 is readily filled out by subtracting one half of the increase from the percentile rank of the top of the interval. Thus the top of the highest interval, or score 90, has a percentile rank of 100, from which we subtract one half of the increase, which for this interval is 1, thus getting 99.5 as the percentile rank of the midpoint. A similar operation gives the percentile rank of the next midpoint below. Since the increase is 2, we subtract $\frac{1}{2} \times 2$, or 1, from 99, the percentile rank of the top of the interval, getting 98 as the percentile rank of the midpoint.

> *Problem:* The student should stop at this point and compute the percentile ranks of the midpoints for the data given in Tables VIII and IX.

Some short methods of computing percentiles. So far we have dealt with the procedure that would be used in the normal course of events, without any effort to introduce short methods which might obscure the essential steps. The following short methods, however, may be of service to some readers by way of reducing the labor involved in making the detailed calculations:

1. Multiplying by the reciprocal of N, instead of dividing by N. Ordinarily, multiplication is shorter and easier than division, and if the reciprocal of N is once computed it may

be multiplied by the various cumulative frequencies, and thus avoid any further division operations. The reciprocal should be carried out to several decimals, however, else the error due to the omitted decimals will be fairly large by the time the hundredth percentile is reached.

2. Computing on the slide rule. If the reciprocal of N is set once for all on the D scale of the slide rule, all the multiplications may be made by simply moving the slide back and forth and thus the routine may be greatly reduced. Also this method avoids the danger of magnifying the error that results from dropping the last decimal places in the reciprocal.

3. Computing on the adding machine. The reciprocal of N may be set on the machine and the key set for cumulative addition, thus making it possible to work directly from the raw frequencies and omitting the column for cumulative frequency. In this case the reciprocal of N is added as many times as there are frequencies in the interval, this subtotal recorded, then the reciprocal is added as many more times as there are frequencies in the next interval, and so on to the top of the table.

The percentile graph. For the visual analysis and presentation of data the percentile graph is extremely effective. Furthermore, it is very easy to construct after the work that has been described up to this point has been completed.

The main essential is to have cross-section paper with ten main vertical lines at equal distances apart. Smaller

divisions may be shown between these, or they may simply be imagined. These ten main vertical lines should be labeled at the top or bottom by writing the numbers 0, 10, 20, etc., up to 100. The scores, frequencies, etc., as shown in Table IX, may then be written on the horizontal lines at the left of the graph so that the lines between the various class intervals extend on across the graph. The percentile for the top of each class interval is then located on that horizontal line by finding the vertical line, or place between two of them, that corresponds to the percentile as represented on the table. The points so located may then simply be joined by a continuous line, and the task is done.

Let us illustrate the construction of the percentile graph by reference to Figure 7. The data are new, but the steps and processes shown by the table at the left are the same

Grammar test scores	*f*	Cum. *f*	Percentile ranks of tops
(1)	(2)	(3)	(4)
85–89	2	54	100
80–84	3	52	96
75–79	7	49	91
70–74	8	42	78
65–69	11	34	63
60–64	6	23	43
55–59	5	17	31
50–54	3	12	22
45–49	3	9	17
40–44	2	6	11
35–39	0	4	7
30–34	1	4	7
25–29	2	3	6
20–24	1	1	2
	54		

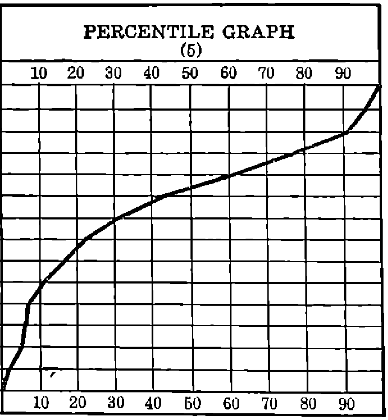


FIG. 7. ILLUSTRATION OF THE CONSTRUCTION OF A PERCENTILE GRAPH

The data are scores of 54 students of college French on a grammar test.

as those in Table IX. Note that the percentiles for the midpoints need not be computed if a graph is to be drawn, since the graph itself will show them without further calculations. The construction of the graph in Figure 7 should begin at the left and the bottom. There are no cases below the line for score 20, hence we start the graph by placing a dot on the vertical zero line where it intersects the horizontal line representing score 20. The percentile for the top of the interval 20–24, or score 24.999, is 2, as shown in column 4, hence we go to the right a distance of 2 units on the horizontal scale and locate a dot. The top of the next interval has a percentile rank of 6, hence we place a dot 6 units to the right of the zero line, and so on until we reach the top and extreme right of the graph, or 100 per cent.

Having drawn the graph, it is easy to find the median, quartiles, or any desired percentiles. To find the median we come down the vertical line for the 50-percentile until it intersects the percentile graph which we have just finished constructing. We then go back to the left horizontally until we reach the column containing the original scores. Since the 50-percentile line intersects the percentile graph in Figure 7 approximately two fifths of the way from the bottom to the top of the class interval for scores 65–69, we may say at once that the median is approximately 67. Similarly, we locate the lower quartile (25-percentile) at approximately the score 57, and the upper quartile (75-percentile) at 74.

Problem: For practice the student should now locate the score for the 20-percentile; the 90-percentile; the 40-percentile.

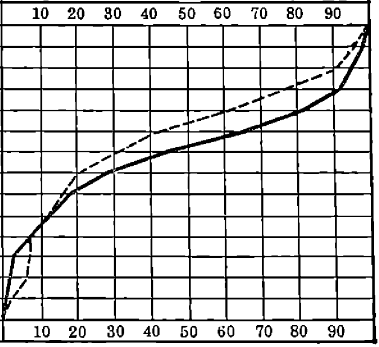
Comparing two distributions by means of percentile graphs. It is now apparent that two or more percentile graphs may be drawn on the same blank, each to represent a different group of data. For example, Table X, page 80, shows the standing of 486 freshmen in a grammar test which was given on entrance to the university. Fifty-four of these took college French, and some one wished to know whether previous ability in grammar had anything to do with the registration for French. Some light may be shed on this question by comparing the 54 French students with the entire 486 students who entered. This is done in Figure 8, which merely brings together in one place the data from Table X and Figure 7.

The reader will note that Figure 8 omits some of the columns which were found in Table X and Figure 7. This is done because the graphs themselves are constructed simply from the columns which show the percentiles, and the inclusion of all the other columns crowds the page unnecessarily. There is no good reason for leaving out these extra columns of computation except mere convenience and economy of labor. They must be computed anyway before the graph is constructed, but need not necessarily be copied in detail.

By looking at Figure 8 it is readily apparent that the students who took college French had entered the university with a very slightly better foundation of grammatical ability, since the dotted line is somewhat higher than the solid line throughout most of its course. We may also compare the median of the two groups by noting the distance be-

Grammar test scores	Percentile ranks of tops: 486 Freshmen	Percentile ranks of tops: 54 French students
(1)	(2)	(3)
85–89	100	100
80–84	99	96
75–79	97	91
70–74	92	78
65–69	82	63
60–64	65	43
55–59	44	31
50–54	29	22
45–49	19	17
40–44	11	11
35–39	7	7
30–34	3	7
25–29	2	6
20–24	1	2

PERCENTILE GRAPH (4)



Key: ——— 486 Freshmen, ---- 54 French students.

FIG. 8. ILLUSTRATION OF THE USE OF PERCENTILE GRAPHS TO COMPARE TWO GROUPS

The data consist of grammar scores earned by 486 college freshmen, compared with those earned by 54 freshmen who took French.

tween the two graphs where they cross the 50-percentile line. Thus the graph for the 486 students intersects it at a point on the level with a score of 62 and the graph for the 54 French students intersects it at a point on the level with score 67. The medians differ, in other words, by about 5 scores, or the amount of one class interval.

Computing percentile ranks from ungrouped data. It is sometimes desirable to compute percentile ranks when there are too few cases to warrant the making of a frequency table. To do so, however, is usually questionable, since the percentiles so computed are very much influenced by chance errors of sampling, as is always the case when generalizations are based on a very few cases. Keeping this

caution in mind, therefore, we may summarize the procedure when computing from ungrouped data in the following statements:

1. The scores should be ranked first, in the order of size, the highest rank being assigned to the largest score.

2. Whenever two or more cases are tied for a given rank they are all assigned a rank in the middle of the range which they would occupy if they had not been tied.

3. The percentile rank for a given case may be found by dividing its simple rank minus .5 by N and multiplying by 100. Thus, if a case is rank 13 in a distribution where N is 19, its percentile rank is $(13 - .5) \times 100/19$, or 66.

Expressing this in a formula we have:

$$\text{Percentile Rank} = \frac{(\text{Rank} - .5) \times 100}{N}$$

The reason for subtracting .5 is that doing so gives the percentile rank of the midpoint, rather than the top of the score in question.

Table XI shows a sample of the process, based on 16 cases. Starting with the smallest score, 4, we find the percentile rank of its midpoint by inserting in the formula, thus getting $(1 - .5) \times 100/16$, or 3. The next score, having rank 2, would yield $(2 - .5) \times 100/16$, or 9. The next three scores, being tied for rank 4, all have a percentile rank of $(4 - .5) \times 100/16$, or 22, and so on throughout the series of scores.

TABLE XI. ILLUSTRATION OF THE COMPUTATION OF PERCENTILES FOR UNGROUPED DATA

(The data are scores on a daily completion test in high-school history.)

Completion test scores	Ranks	Percentile ranks
(1)	(2)	(3)
22	16	97
21	14.5	88
21	14.5	88
20	13	78
18	12	72
17	11	66
15	10	59
14	9	53
13	7.5	44
13	7.5	44
10	6	34
7	4	22
7	4	22
7	4	22
6	2	9
4	1	3

TRUE-FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. The best student in the group should always be rank one.
2. The percentile rank of a student indicates exactly what per cent of the students in his own class surpassed him.
3. The total for the cumulative-frequency column should always be equal to N.
4. Percentile graphs are smoother than frequency polygons.
5. The forty-percentile point divides the upper two fifths from the lower three fifths of a group.
6. The percentile rank of a pupil indicates the rank he would have if there were exactly a hundred pupils in the group.

7. Percentile ranks are rather unreliable if based on very small numbers of cases.
8. Percentiles may be classed as a common denominator in dealing with rank data.
9. The percentile graph should be plotted from the cumulative frequency column.
10. Two groups are easier to compare on the basis of percentile graphs than frequency curves, because there is less crossing of lines.
11. Percentiles and percentile ranks mean exactly the same thing.
12. If there were exactly one hundred people in a group, the percentile rank of the best student should be 99.5.
13. The median is the fiftieth percentile.
14. In plotting the percentile graph the upper limit of each class interval is located, rather than the midpoint.
15. The same principle is used in finding percentiles as in finding medians and quartiles.
16. A percentile is a point and not a distance.
17. The percentile rank of the midpoint is found by adding one half of the increase to the cumulative frequency.
18. The median is easily found from the percentile graph.
19. It is necessary to find the percentile ranks of midpoints in order to draw the percentile curve.
20. All percentile curves tend to have the same characteristic shape or form.

QUESTIONS AND EXERCISES

1. Below are scores earned by two classes on a true-false test containing 20 statements. The first class had studied the material while the second had not. Find the percentile ranks for the scores in each class, and compare the two groups by constructing their graphs on the same page. What do the two curves show about the value of the test as a measuring device?

Test scores (No. right)	Frequency	
	Class I	Class II
(1)	(2)	(3)
20	1	
19	3	
18	8	
17	19	
16	57	2
15	51	4
14	28	13
13	17	12
12	3	8
11		7
10		4
9		2
Total	187	52

2. From the graphs locate the medians and quartiles.
3. How many scores better is the median of Class I than Class II?
4. What per cent of Class II exceeded the lower quartile for Class I? The median of Class I?
5. Compute the percentiles and draw the percentile curves for all sets of data given in exercises 7, 8, and 9 at the end of Chapter III.
6. Compute the percentiles and draw the percentile curves for the data given in the exercises at the end of Chapter IV, using tabulations in intervals of 5 and also in intervals of 8.

CHAPTER VII

VARIABILITY

The meaning of the term. There is constant and frequent need for setting limits to our generalizations. Thus we may say that the children in a given grade are "12 years old, more or less." We mean by this that the average age is about 12 years, but that the children vary above and below this central tendency. A statistician would not be content with the statement "more or less," but would want to know *how much* more or less. He would learn possibly that the youngest child was nine and the oldest sixteen, and that half of the children were either 11, 12, or 13 years old. This would give him a very definite idea of the ages represented by the group. Such a statement is simply a means of describing the *variability* of a group.

By variability we mean the amount of scatter or the amount of spreading out of cases from the central tendency, or average. The variability is small if the scores are closely concentrated within a small range, and large if the scores spread out so that the lowest and the highest are far apart.

Cases and situations in which measures of variability are valuable. The value and importance of definite measures of variability may be illustrated in terms of a few concrete school situations, as follows:

1. A principal wishes to know how well he has his school classified into grades, so that pupils who are grouped together are of equal ability and are able to work together

effectively. He gives standardized tests in reading, arithmetic, etc., and notes the spread or variability of scores in each grade. If this variability is very great, it indicates that pupils are poorly classified, and that some pupils of high ability and others of low ability are grouped together. If, on the other hand, each grade shows a very low spread, range, or variability, the conclusion is that each grade is quite homogeneous and that those who are classified together are well able to work together.

2. A test had been given to two classes, and one class had earned an average of 5 points more on it than the other. It was desired to know how much significance to attach to a difference of 5 points. This question could be answered in part by finding how widely the scores varied. Thus if class A had individual-pupil scores ranging between 35 and 40, and class B between 41 and 45, with no overlapping at all, the difference would be very significant, since the poorest student in class B exceeded the best one in class A. If, on the other hand, the scores in each class ranged from 10 to 70, and the mean scores were 5 points apart, this small difference would be of little significance. In other words, a difference between two averages is not very meaningful until we learn the variability of the two sets of scores.

The point mentioned above may be made clear by reference to Figure 9. The difference between the means in each pair is the same, but this difference is very important in case 1 because the variability is low, and not very important in case 2 because the variability is much higher, resulting in more overlapping of the two curves.

3. A teacher gave two tests to her class. One contained 40 true-false items and the other 40 completion items. She wished to combine the scores on the two tests so that each type of test would be equally influential in determining the final grades of the pupils in the class. She found, however, that whereas the true-false test scores ranged between 20 right and 40 right, the completion test scores were more widely scattered and ranged between 0 and 40. In other words, the completion test scores had twice as great variability as the true-false scores. This fact would therefore cause the completion test to influence the composite score twice as much as the true-false test, whereas she wished to weight them equally.

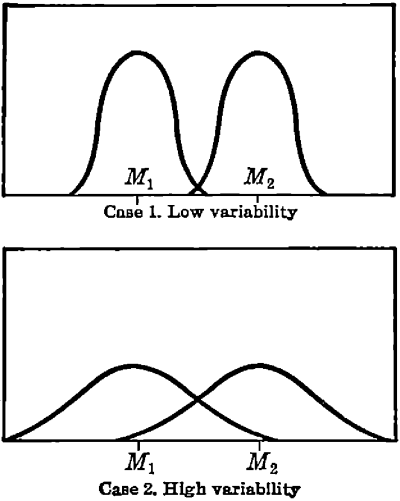


FIG. 9. ILLUSTRATION OF THE NEED OF TAKING ACCOUNT OF VARIABILITY WHEN COMPARING TWO DISTRIBUTIONS

The data are purely imaginary, and the curves are drawn smooth for greater simplicity.

The solution was simple. She multiplied the true-false scores by 2, thus doubling their variability, after which the two tests could be combined with the assurance that each would have equal influence on the composite grade. If she had wanted the true-false test to count twice as much as the

completion test, she would have multiplied the true-false scores by 4. Similarly, any two sets of scores can be weighted in any desired proportions if their respective variabilities are determined, and one set of scores multiplied by a number which will yield the desired ratio between the variabilities of the two sets of scores.

Different methods of measuring or describing variability. There are several different methods by which statisticians indicate the amount of variability of data, including the range, the frequency polygon, the percentile curve, selected percentile points, quartile deviations, standard deviation, and others. These will be described individually.

The range. A very simple, but not very helpful, measure of variability is the range, or the difference between the lowest and highest scores. For some purposes it is adequate, but ordinarily it is of small value because it is affected to a great extent by the one or two very exceptional cases that fall at the ends of the range.

The frequency polygon. A very effective, but a somewhat cumbersome way to represent the variability of a distribution is to draw a frequency polygon describing the entire set of data. This gives a complete visual picture of the entire distribution, showing how far above and below the mean the cases vary, and how flat or steep the curve is. Statisticians use the frequency polygon very extensively to serve this purpose.

The percentile curve. The percentile curve serves somewhat the same purpose in describing variability as does the frequency polygon. This will be clear if the reader will

turn back to the previous chapter. It has the same limitations as the frequency polygon in that it is cumbersome to handle, and does not lend itself readily to translation into verbal statements.

Quintiles, deciles, and other selected percentiles. To eliminate some of the detail and cumbersomeness of percentile curves and frequency polygons, it is quite common to choose a few handy points and leave out of account the rest. Thus one might say the median of a distribution was 72, that the 90-percentile was 113, and the 10-percentile was 43. Or he might say the 80-percentile, or fourth quintile, was 101, and the 20-percentile, or first quintile, was 56. This gives a fairly good idea as to the range, but leaves out of account the 10 or 20 per cent of scores that fall at the extreme ends of the curve.

Quartiles and quartile deviation. When selected percentile points are used to describe variability, as suggested in the previous paragraph, it is more common to select the quartiles, or 25- and 75-percentiles, as these points. Thus if we report the median and these two quartile points we give a fairly adequate description of the distribution, since we set forth the limits within which the middle 50 per cent of the cases fall.

A distinction should be made between the term "quartiles" and the term "quartile deviation." The quartiles are points, and the quartile deviation is *one half the distance between them.* It is customary to consider the quartile deviation as one half the distance between the quartiles, rather than the whole distance, since one half of the total

distance falls on either side of the median (in the normal curve). The total distance between the quartile points is sometimes spoken of as the *interquartile range*, and the quartile deviation is then called the *semi-interquartile range.*

Calculation of quartile deviation. The computation of the quartile deviation is very simple. It consists of subtracting the lower quartile from the upper quartile and dividing by 2. It may be expressed in a formula, as follows:

$$Q = \frac{Q_3 - Q_1}{2}$$

The only difficulty here lies in finding Q_1 and Q_3. This is done in the same general manner as finding the median, which in reality is Q_2. To find Q_1 count up from the bottom until 25 per cent of the cases have been passed. If this point falls inside of an interval, it is necessary to interpolate, as in finding the median. (See page 65.) The same process may be applied to find Q_3, simply counting up until 75 per cent of the cases have been passed and interpolating in the last interval, if necessary. The process is illustrated in Table XII.

Standard deviation (σ). We have just studied quartile deviation, which measures the middle half of the distribution. We shall now consider a different measure, the *standard deviation*, represented by the Greek letter σ, which measures approximately the middle two thirds of the distribution (68.26 per cent, to be exact). If, then, we describe a distribution by saying it has a mean of 65 and a σ of 10, we mean that if we begin with score 55 and include everything up to score 75, we include two thirds of the cases of

TABLE XII. ILLUSTRATION OF THE COMPUTATION OF QUARTILE DEVIATION

(The data are purely imaginary.)

Scores (1)	f (2)
80–84	1
75–79	3
70–74	15
65–69	10
60–64	13
55–59	21
50–54	28
45–49	22
40–44	19
35–39	13
30–34	9
25–29	4
20–24	2
Total	160

Q_1 is at the $1/4 \times 160$ frequency, or at the 40th frequency, or in the 40–44 interval.

Therefore $Q_1 = 40 + \frac{12}{19} \times 5 = 43.2$

Likewise, Q_3 is at the $3/4 \times 160$ frequency, or at the 120th frequency, or in the 60–64 interval.

Therefore $Q_3 = 60 + \frac{2}{13} \times 5 = 60.8$

Similarly, the median, or Q_2, is at the $1/2 \times 160$ frequency, or at the 80th frequency, or in the 50–54 interval.

Therefore $Q_2 = 50 + \frac{11}{28} \times 5 = 51.9$

Formula:

$$Q = \frac{Q_3 - Q_1}{2}$$

$$= \frac{60.8 - 43.2}{2}$$

$$= 8.8$$

(*Note:* It is not necessary to find the median in order to get the quartile deviation, but it has been done here to show how it resembles quartiles as to methods of computation.)

the distribution. The chances are therefore one in six that a given case will be above 75, and one in six that it will be below 55. In the normal curve the σ distance would reach down approximately to the 16th percentile and up to the 84th percentile, but if the distribution is irregular or skewed, these percentiles may be somewhat different, since the σ is influenced by the size as well as the number of extreme variations.

In finding the σ of a distribution, it is possible to begin the computation from the true mean, but this is rarely done because it ordinarily involves the squaring of fractions. When the true mean is used, the procedure consists of dividing the sum of the squares of the deviations by N and extracting the square root of the resulting quotient

However, the plan of computing deviations from an assumed mean and then making a correction as is done in computing the mean (see Chapter IV) is commonly employed. This involves very little additional work after the mean has been computed since the σ uses most of the previous work and merely adds one more step. We shall illustrate the process in Table XIII. Columns 1–4 are familiar to us (see chapter on the mean). Column 5, fd^2, is obtained by multiplying the entries in columns 3 and 4. The total, Σfd^2, is simply the result of simple addition. All numbers in column 5 are positive because the product of two negative numbers is always positive. The formula as given below the table requires merely the substitution of the various items and then clearing up algebraically.

The principal danger of error lies in forgetting to multiply

TABLE XIII. ILLUSTRATION OF THE COMPUTATION OF THE STANDARD DEVIATION

(The data are scores of college freshmen on a reasoning test.)

Reasoning test scores	f	d	fd	fd^2 (product, columns 3 × 4)
(1)	(2)	(3)	(4)	(5)
18–19	2	5	10	50
16–17	7	4	28	112
14–15	10	3	30	90
12–13	22	2	44	88
10–11	29	1	29	29
8– 9	40	0	0	0
6– 7	26	−1	−26	26
4– 5	16	−2	−32	64
2– 3	9	−3	−27	81
0– 1	3	−4	−12	48
Totals	164 (N)		44 (Σfd)	588 (Σfd^2)

$$M = 9 + 44/164 \times 2 = 9.54$$

Formula:[1]

$$\sigma = \left[\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}\right] \times \text{Interval}$$

$$\sigma = \left[\sqrt{\frac{588}{164} - \left(\frac{44}{164}\right)^2}\right] \times 2$$

$$= 2\sqrt{3.52}$$

$$= 3.76$$

by the class interval, thus making σ very much too small. A good way to check up on this and other major errors is to inspect the data to see if two thirds of the cases fall within the limits of one σ above and below the mean. Thus in our own case we locate the upper point at 9.54 + 3.76, or 13.30,

[1] For derivation of this formula, see Kelley, T. L., *Statistical Method*, pp. 77–82.

and the lower point at 9.54 − 3.76, or 5.78. If our work is accurate we may expect two thirds of the 164 cases, or about 110 cases, to fall between these limits. The point 13.30 takes about the bottom two thirds of the 22 cases in the 12–13 interval, or about 15 cases. The point 5.78 falls slightly below the top line of the 4–5 interval, theoretically taking about one eighth of the 16 cases in this interval, or 2 cases. We then add the cases between these two points as follows: $15 + 29 + 40 + 26 + 2 = 112$. We know, therefore, that the work is correct, except for possible minor errors.

A model computation of the two principal measures of variability. In order further to explain and illustrate the quartile deviation and standard deviation, Table XIV is included and both of these quantities have been computed from the same original data. The reader should check the work carefully until he understands how each of the items has been computed, after which this table may be used as a model in dealing with other distributions.

Comparing σ with Q, we note that it is approximately half again as large, or, to say it differently, Q is about two thirds the size of σ, as we should expect.

Average or mean deviation. A measure of variability which is occasionally used is the average or mean deviation, expressed as *A.D.* or *M.D.* It is found by adding together all deviations from the mean regardless of sign and dividing by N. Expressed in a formula it would be as follows:

Formula:[1]
$$A.D. = \frac{\Sigma fd'}{N}$$

[1] For derivation of the formula and a more extensive treatment of *A.D.*, see Holzinger, K. J., *Statistical Methods for Students in Education*, pp. 102–107.

TABLE XIV. ILLUSTRATION OF THE COMPUTATION OF QUARTILE AND STANDARD DEVIATION

(The data are from Table II, Chapter III.)

Scores	f	d	fd	fd^2
(1)	(2)	(3)	(4)	(5)
135–139	1	5	5	25
130–134	1	4	4	16
125–129	3	3	9	27
120–124	2	2	4	8
115–119	6	1	6	6
110–114	15	0	0	0
105–109	9	−1	−9	9
100–104	4	−2	−8	16
95–99	7	−3	−21	63
90–94	2	−4	−8	32
85–89	3	−5	−15	75
80–84	2	−6	−12	72
Totals	55 (N)		−45 (Σfd)	349 (Σfd^2)

$$M = 112.5 + (-45/55 \times 5) = 108.4$$

Formula:

$$Q = \frac{Q_3 - Q_1}{2}$$

Q_1 is at the $\frac{1}{4} \times 55$ frequency or the $13\frac{3}{4}$ frequency, or in the 95–99 interval

$$Q_1 = 95 + \left(\frac{6.75}{7} \times 5\right) = 99.82$$

Q_3 is at the $\frac{3}{4} \times 55$, or $41\frac{1}{4}$ frequency, or in the 110–114 interval

$$Q_3 = 110 + \left(\frac{14.25}{15} \times 5\right) = 114.75$$

$$Q = \frac{114.75 - 99.82}{2} = 7.47$$

Formula:

$$\sigma = \left[\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}\right] \times \text{Int.}$$

$$= \left[\sqrt{\frac{349}{55} - \left(-\frac{45}{55}\right)^2}\right] \times 5$$

$$= 11.9$$

The reader should note that d' is used instead of d, which has been familiar in various computations at other points in this book. The d' here stands for deviations from the true mean, whereas d has always been used for deviations from the assumed mean. The procedure would be to compute the mean in the usual way and then count deviations from this instead of the assumed mean, which will naturally give fractional deviations rather than whole numbers. Deviations should be counted in terms of actual scores, rather than class intervals, to avoid confusion. After this is done the deviations are multiplied by the frequencies for the various class intervals, and the total of these fd' items, or Σfd, is divided by N.

It would be possible to work from an assumed mean, using a correction, but doing so involves certain technical difficulties which make it unwise to attempt to present the method here. In fact the authors of this book feel that the elementary student of statistics should concentrate his energies on a few standard methods of doing a piece of work, rather than scatter over many and learning them poorly. For that reason they have refrained from including a model computation for the average deviation, and would suggest that students not try to learn its actual computation. The above general description will serve to provide an understanding of the mean deviation if it is met in reading educational literature, and the writers feel that it is unwise to attempt more than this in an elementary course.

When to use each measure of variability. Of the three principal mathematical ways of describing variability,

which is preferable? For what is each best suited? The answer is not to be taken dogmatically and arbitrarily, but in general it is as follows:

1. Use quartile deviation when the median is used to describe the central tendency.

2. Use standard deviation when the mean is used for central tendency.

3. Do not use average or mean deviation at all in the beginning stages of statistical work. It is used by certain writers and statistical workers, but its purpose can be served fairly well by quartile deviation or standard deviation. Since it is used in educational literature it is well to know what it means, but the authors of this book recommend that the beginner concentrate on the other measures of variability and not attempt to use average deviation in actual computations. In advanced work problems will be faced which require its use, and the computation of it can be learned at that time.

TRUE-FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. Variability is relatively unimportant in classifying students for instructional purposes.
2. Quartile deviation is affected less by a skewed distribution than is standard deviation.
3. The standard deviation is the perpendicular line erected where the curve changes from concave to convex.
4. Quartile deviation is approximately the distance between the first quartile and the median.

5. Much of the work done in computing the mean may be utilized in the computation of the standard deviation.
6. Variability of scores means the amount of scatter or spread away from the central tendency.
7. In combining scores from several tests, the test that shows the highest variability gets highest weight in the composite.
8. The greater the variability, the more accurate is our knowledge of a group.
9. The method of computing the quartile deviation involves counting cases in the same manner as in finding the median.
10. The standard deviation is a distance and not a point.
11. Sectioning students according to ability reduces the variability of student work within a given class, but not within a given school.
12. The quartile deviation is the most accurate unit of variability.
13. One standard deviation marked off on each side of the mean in a normal curve takes in 68.26 per cent of the distribution.
14. Variability supplements the average as an aid in describing a distribution.
15. Quartile deviation is the distance between the first and third quartiles.
16. The range is the most perfect measure of variability because it includes all the cases.
17. Computation is shortened if the standard deviation is computed from the true mean rather than an assumed one.
18. It is necessary to compute the mean before you can get the standard deviation.
19. In finding the standard deviation the deviations are squared.
20. Average deviation is more important than standard deviation for measuring variability.
21. Quartiles and quartile deviations mean the same thing.

QUESTIONS AND EXERCISES

1. Define variability or dispersion.
2. Explain how two distributions may have the same range and average, and yet differ in variability.
3. Give the steps in computing Q.
4. Give the steps in finding the standard deviation.
5. Compute Q and σ for the distributions in:
   a. Figure 4, Chapter III (data for Test 1).
   b. Figure 5, Chapter III (data for boys; also for girls).
   c. Table VIII, Chapter VI.
   d. Table IX, Chapter VI.
   e. Table X, Chapter VI.
   f. Figure 7, Chapter VI.
6. In problem 5, plot the frequency polygon and indicate, as closely as possible, the points representing the σ distances, and erect lines perpendicular to the base at these points.

CHAPTER VIII

THE NORMAL CURVE

The normal-frequency curve. Every teacher is aware of the fact that pupils differ greatly in ability. Some pupils, on the basis of such measurements as we can make, have scores three to five times as high, numerically, as other pupils in the same class. Yet, if we measure any considerable group of pupils in any ability or trait, where we have relatively accurate measurement devices, we are impressed with the fact that the resulting distributions always tend to be symmetrical, or bell-shaped. This is true whether we measure height, intelligence, or ability in the fundamentals in arithmetic.

This symmetrical or bell-shaped distribution is so nearly universal in statistics that it has come to be called the *normal curve.* It is known by a number of other names, among them the biological curve, the normal-frequency curve, the normal-probability curve, and the binomial curve. Each of these names is related to a type of data from which evidence of the existence of such a curve is obtained. Regardless of the particular name used, however, what is meant is a certain constant arrangement of data about a central tendency which can be expressed in exact mathematical language, and which forms the basis for important computations in statistics.

The normal curve and chance. If we tabulate and plot the combinations of heads and tails obtained by tossing a

number of coins a great many times, we are impressed by the extent to which the resulting curve approximates the normal curve. In numerous such experiments reported by statisticians, involving throwing dice, drawing numbers out of a basket, etc., curves plotted from the resulting combinations of results were found consistently to approximate very closely the normal curve.

In each of these cases, the resulting combinations are theoretically due to pure chance. Thus, if one coin is tossed, there is a 50–50 chance of having a head or tail turn up. However, as soon as more than one coin is tossed, the possible combinations become greater. For example, if two coins are tossed, there are four possible ways in which they may fall (as illustrated in Table XV) giving the following combinations: two heads, two tails, a tail and a head, and a head and a tail. Since the latter two consist of a head and tail each, they may be considered as two opportunities to get a head-tail combination. Of the four possible combinations, therefore, when two coins are tossed, the chances of obtaining each combination are as follows:

TABLE XV. ILLUSTRATION OF THE POSSIBLE COMBINATIONS OF HEADS AND TAILS WHEN TWO COINS ARE TOSSED

Possibilities	First coin	Second coin
1	H	H
2	H	T
3	T	H
4	T	T

Two tails: one out of four (25%)
One head and one tail: two out of four (50%)
Two heads: one out of four (25 %)

A still greater number of combinations are possible if more than two coins are used. Table XVI illustrates the eight combinations possible when three coins are used.

TABLE XVI. ILLUSTRATION OF THE POSSIBLE COMBINATIONS OF HEADS AND TAILS WHEN THREE COINS ARE TOSSED

Possibilities	First Coin	Second Coin	Third Coin
1	H	H	H
2	H	T	T
3	H	T	H
4	H	H	T
5	T	T	T
6	T	H	H
7	T	H	T
8	T	T	H

An inspection of the figure reveals that the chances of obtaining each combination are as follows:

3 heads: One out of eight ($12\frac{1}{2}\%$)
2 heads, 1 tail: Three out of eight ($37\frac{1}{2}\%$)
1 head, 2 tails: Three out of eight ($37\frac{1}{2}\%$)
3 tails: One out of eight ($12\frac{1}{2}\%$)

While for only a few tosses the number of combinations of each kind might not agree exactly with what would be expected theoretically, in a large number of tosses results should agree closely with the percentages given.

A similar table could be prepared showing the possible combinations and their expected frequencies in cases where four, five, six, or more coins were tossed. It is needless to carry the process out here in concrete form, however, since the above two examples illustrate the principle involved. We shall, therefore, go on to the mathematical statement of the whole matter, as it is involved in the binomial expansion.

The normal curve and mathematics. We have seen that the data derived from purely chance situations provide experimental evidence in favor of the normal-curve hypothesis. The field of mathematics, however, and specifically the binomial theorem, gives the exact theoretical explanation of these observed facts, and makes clear the general law that runs through all such experiments as coin tossing or dice throwing.

In the discussion of coin tossing we illustrated the theoretical possibilities (Table XVI) when three coins were used. Instead of listing all of the possibilities as in Table XVI, we could have attained the same end by reducing our problem to the form of the binomial $(H + T)^3$, where H represents the chance of obtaining a head when 3 coins are tossed, i.e., when the power to which the binomial is to be raised is 3. Expanding the binomial[1] $(H + T)^3$ gives us

$$H^3 + 3\ H^2T + 3\ HT^2 + T^3$$

[1] For the benefit of those who do not remember how to expand a binomial, the rule is as follows: The first term of the expansion is the first term of the binomial with an exponent equal to the power to which the binomial is to be raised; thereafter, each succeeding coefficient in the expansion (coefficient of the first term is one) is equal to the coefficient of the preceding term, multiplied by the exponent of the first letter of the preceding term, divided by the number of the preceding term. Thus the coefficient for the second term of $(H + T)^3$ is equal to 3, because in the first term of the expansion (H^3) the coefficient 1 (understood) multiplied by the exponent 3 and divided by 1 (the number of the term) is equal to 3. The second term of the binomial is introduced into the second term of the expansion with an exponent of 1 understood. Thereafter the exponents of the first term of the binomial become less each time by 1, and the exponents of the second term increase by one each time until equal to the power to which the binomial is to be raised. The same result will be secured if the multiplication is carried out in detail, thus: $(H + T)(H + T)(H + T)$.

For a more detailed treatment of the binomial theorem and the normal curve, see Yule, G. U., *An Introduction to the Theory of Statistics*, chap. xv, or Thurstone, L. L., *The Fundamentals of Statistics*, chap. xvii and xviii.

The coefficients of H^3 and T^3 are, of course, 1. Adding all the coefficients gives us the sum of eight possibilities listed in Table XVI. 1 H^3 is to be interpreted as one chance to get 3 heads; 3 H^2T, as 3 chances to get 2 heads and 1 tail, etc.

If instead of tossing 3 coins we use 4, our binomial becomes $(H + T)^4$, and its expansion gives us

$$H^4 + 4\ H^3T + 6\ H^2T^2 + 4\ HT^3 + T^4.$$

If we plot this expanded binomial, using coefficients as frequencies, our result will appear as in Figure 10. The

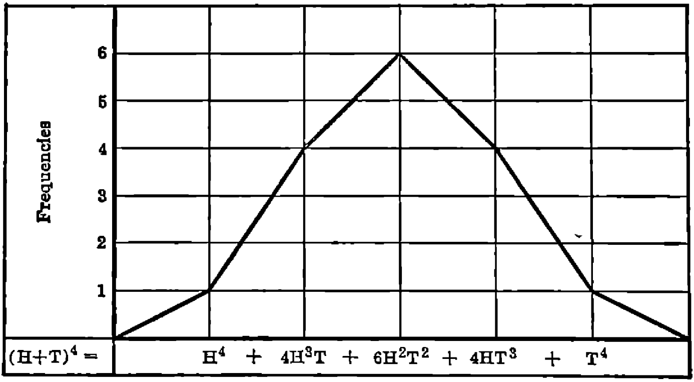


FIG. 10. ILLUSTRATION OF THE CURVE OBTAINED FROM PLOTTING THE EXPANDED BINOMIAL

larger the power to which we expand a binomial the more points will be represented on the curve, and the more closely it will resemble the conventional smooth bell-shaped curve.

Why educational and biological data approximate the mathematically derived curve. We have seen that the

mathematical expansion of the binomial is simply a precise formulation of what happens when combinations are made according to pure chance. What right have we, however, to assume that it applies to biological and educational data? A little study of the biological basis of heredity will make this clear. The biologists tell us that the traits of the offspring are determined by the combinations of chromosomes from the male and the female germ cells, and that such items as sex, eye color, height, or intelligence depend upon the particular combinations of chromosomes that take place. Intelligence, for example, may be assumed to be due to a large number of different contributing factors; and one child may be so fortunate as to "draw a high card" on each of these, thus making him a genius; while another, not so fortunate in the lottery, instead of drawing "all heads" gets a combination involving very few of the favorable elements. This explanation will show in a general way that educational and biological data may really be interpreted as applying the laws of chance just as truly as do such phenomena as drawing cards, tossing coins, or throwing dice.

Additional mathematical properties of the normal curve. We have now considered the normal curve in its relation to the element of chance and to the expansion of a binomial; and we have considered the explanation given by biologists for the close agreement of educational and biological curves with the mathematically derived curve. There are, in addition, some important specific aspects of this curve that will next receive attention.

As may be inferred from previous discussions, one of the important characteristics of the normal curve is its symmetry, the fact that the slope in either direction from its highest point is exactly proportional, and at exactly the same distance on each side of the midpoint the curve changes from convex to concave.

Another aspect of the normal curve which deserves some attention is the manner in which it approaches the base line. For practical purposes, this curve is often broken off at 2.5 or 3 σ in either direction from the mean, giving the impression that it touches the base line at these points. This, however, is not the case. While in a particular restricted sample all data may fall within these limits, if all possible data of this type were available, some cases would, theoretically, always fall outside these limits. It is for this reason that, in drawing the curve, it is never quite allowed to touch the base line, but extends indefinitely, approaching nearer and nearer to it.

Curve areas and standard deviation. In the chapter on variability we studied the σ and learned how to compute it. We discovered that if one σ unit is measured off in either direction ($\pm\ \sigma$) from the mean of a distribution, these limits will enclose the middle two thirds (68.26 per cent) of the distribution, taking into account both the number and size of the scores or measurements that have been tabulated.

It is frequently desirable to know what part of a distribution is enclosed by more or less than exactly 1 σ. The fact that, in a normal distribution (by definition) these areas

of the curve bear a constant relationship to the σ makes it possible to prepare tables giving the exact percentages of the area of a curve included between perpendiculars erected at various distances from the mean, such as 1.1 σ, 2.3 σ, 3.7 σ, or any other fractional amount.

There is one important fact in this connection which must be borne in mind. One σ on the base line of a polygon means a certain definite distance in terms of scores or other measures. The second or third σ (laid off in either direction from the mean) on the base line is exactly the same length as the first. But very obviously, due to the fact that the sides of the curve gradually slope toward the base on either side of the mean, the area of a distribution taken in by a second σ will be considerably less than that taken in by the first; and the third σ will take in less than the second. Thus in the normal curve illustrated in Figure 11, it may be seen that the first, second, and third σ's take in 34.13, 13.59, and 2.15 per cent of the distribution respectively.

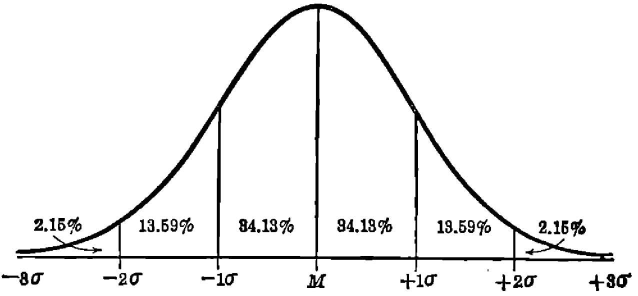


FIG. 11. ILLUSTRATION OF THE PERCENTAGES OF THE TOTAL DISTRIBUTION CORRESPONDING TO SUCCESSIVE σ UNITS FROM THE MEAN

Note that the total of the percentages given here is only 99.74 per cent, instead of 100 per cent. The difference represents the insignificant portion of the curve which extends beyond 3 σ units.

From the foregoing discussion, it is apparent that if the parts of a distribution corresponding to each σ distance become smaller as we go farther from the mean, the parts of a distribution corresponding to succeeding fractional parts of a σ must also become smaller. Thus the first .1 σ takes in 3.98 per cent of the distribution, but the first tenth of the second σ takes in only 2.30 per cent. No two equal fractional parts of a σ on the same side of the mean take in identical amounts of the distribution because each succeeding unit of area becomes smaller because of the slope of the curve. For this reason it is necessary to use the mean, rather than the bottom or top of the distribution, as the starting point in making computations based on the mathematical properties of the normal curve.

In general, therefore, it is possible to make the three following statements about the mathematical properties of the normal curve:

1. The per cent of a distribution corresponding to any σ value in either direction from the mean is constant; that is, the percentage of measurements, scores, or cases between the mean and any σ value from the mean is always the same. In other words, the curve is symmetrical about its mean

2. When we know in terms of σ units how far a point is from the mean, we can tell what per cent of the distribution lies between that point and the mean.

3. If we know what per cent of the scores, cases, or measurements (per cent of distribution) lies between the

mean and a given point, we can find the number of σ units which that point is distant from the mean.

The σ table. Table XVII is known as a sigma or σ table. It gives the percentage of cases (per cent of a normal distribution) which lie between the mean and each σ unit value

TABLE XVII. PERCENTAGES OF THE NORMAL DISTRIBUTION LYING BETWEEN THE MEAN AND VARIOUS σ UNIT VALUES

σ units	Per cent of cases	σ units	Per cent of cases
.00	0.00	1.50	43.32
.05	1.99	1.55	43.94
.10	3.98	1.60	44.52
.15	5.96	1.65	45.05
.20	7.93	1.70	45.54
.25	9.87	1.75	45.99
.30	11.79	1.80	46.41
.35	13.68	1.85	46.78
.40	15.54	1.90	47.13
.45	17.36	1.95	47.44
.50	19.15	2.00	47.72
.55	20.88	2.05	47.98
.60	22.57	2.10	48.21
.65	24.22	2.15	48.42
.70	25.80	2.20	48.61
.75	27.34	2.25	48.78
.80	28.81	2.30	48.93
.85	30.23	2.35	49.06
.90	31.59	2.40	49.18
.95	32.90	2.45	49.29
1.00	34.13	2.50	49.38
1.05	35.31	2.55	49.46
1.10	36.43	2.60	49.53
1.15	37.49	2.65	49.60
1.20	38.49	2.70	49.65
1.25	39.44	2.75	49.70
1.30	40.32	2.80	49.74
1.35	41.15	2.85	49.78
1.40	41.92	2.90	49.81
1.45	42.65	2.95	49.84
		3.00	49.87

up to 3 σ. Other tables have been developed for special purposes extending to 5 σ units from the mean (see Table XIX), but since $\pm$ 3 σ encloses more than 99 per cent of the normal distribution, a 3 σ table is sufficient for most practical purposes.

Its use is a very simple matter. If we wish to know how many cases lie between the mean and a point 1.4 σ from the mean, we find 1.4 σ and read the result 41.92 (41.92 per cent or 4192 cases in 10,000). If the perpendiculars are erected to the base line of a polygon at the mean and at a distance of 1.4 σ from the mean, the area enclosed will contain 41.92 per cent of the area of the entire distribution. If we know that 17.36 per cent of a distribution lies between a perpendicular erected at a certain point and a perpendicular erected at the

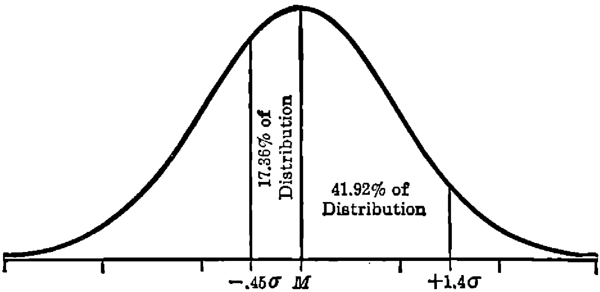


Fig. 12. Illustration of the Relation Between σ Unit Values and Fractional Parts of a Normal Distribution

mean, and we wish to know the distance of this point from the mean in terms of σ, we look for this per cent value in the table, and discover that its equivalent is .45 σ. These two computations are illustrated in Figure 12.

Problem: At this point the student should stop and practice the use of Table XVII. For example:

1. Find what percentage of the area lies between the mean and a point 1.95 σ from it; 2.6 σ; 2.35 σ; .65 σ.

2. What σ values correspond to the following per cents of the distribution: 32.12 per cent; 44.18; 49.22 per cent?

In some of these cases, it will be necessary to interpolate, or else simply to use approximations.

The normal curve and the distribution of teachers' marks. There has been much discussion in recent years regarding the use of the normal curve as a basis for assigning marks to students. Several different percentage bases for the distribution of marks have been proposed, each being recommended as based on the normal curve. Thus we have such diversity as is shown in Table XVIII, and still each system can be definitely shown to be in harmony with the normal curve. The differences are due to the fact that their various advocates have divided the base line of the curve at different points or in different proportions.

TABLE XVIII. PERCENTAGES OF STUDENTS ASSIGNED EACH MARK UNDER VARIOUS SYSTEMS THAT USE THE NORMAL CURVE AS A BASIS FOR MARKING

System	Per cent of students for each mark				
	E	D	C	B	A
(1)	(2)	(3)	(4)	(5)	(6)
1	7	24	38	24	7
2	2	23	50	23	2
3	3	22	50	22	3
4	4	24	44	24	4
5	10	15	50	15	10
6	10	20	40	20	10

Among the schemes of uniform grading that are commonly employed, that which involves 7 per cent E's, 24 per cent D's, 38 per cent C's, 24 per cent B's, and 7 per cent A's seems to be preferable in many ways. These percentages (approximately) are derived from dividing the base line of the normal curve into five σ units, 2.5 of which are on either side of the mean, and throwing the few cases that fall further out than 2.5 σ units in with their adjacent subdivisions.

Thus the grade of C is given to the students in the middle section of the curve, ranging between $+.5\ \sigma$ and $-.5\ \sigma$; B is given to the cases from $+.5\ \sigma$ up to $+1.5\ \sigma$; and A is given to those above $+1.5\ \sigma$, even if a few of these are more than $+2.5\ \sigma$. Likewise, D is given to the students who are between $-.5\ \sigma$ and $-1.5\ \sigma$; and E, or failure, is given to all cases below $-1.5\ \sigma$. Table XVII will show that when these points are used as subdivisions, the percentages of the total distribution are approximately 7, 24, 38, 24, and 7.

This distribution will be more readily apparent from Figure 13, which describes graphically the basis of this system of marking. A comparison of Figure 13 with Figure 11 will be helpful in calling attention to the fact that the divisions are made each time not at whole σ distances, but at half σ distances. Thus the division points in Figure 11 are 1 σ, 2 σ, and 3 σ, whereas in Figure 13, they are at .5 σ, 1.5 σ, and 2.5 σ.

We should make clear here the fact that the principal argument for applying the normal curve in the distribution of grades is not that one percentage distribution is inher-

ently better than some other, but rather that *any definitely-agreed-upon percentage* is preferable to the anarchy that results when each teacher is his own and only judge as to how many high or low grades he will give.

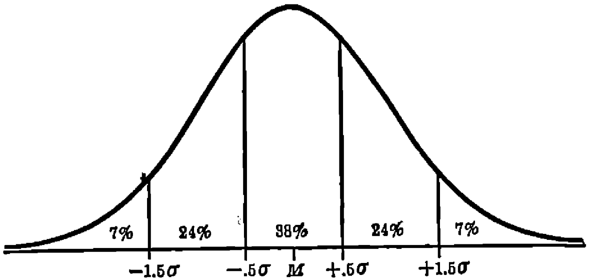


FIG. 13. ILLUSTRATION OF THE METHOD OF DIVIDING THE NORMAL DISTRIBUTION CURVE INTO FIVE PARTS FOR THE ASSIGNMENT OF TEACHERS' MARKS

Determining the difficulty of examinations or test questions. The teacher of a generation ago often was greatly worried because the examination papers of her classes sometimes showed many marks of 40 to 60 per cent, whereas 75 per cent was necessary for passing. She often ascribed this to lack of ability on the part of the class, or to poor teaching on her part, or both. And if she raised some of these marks so that only a few students failed, she undoubtedly felt very guilty and probably kept the fact a secret.

The well-trained teacher of to-day has no such worry, because she knows that the old device of 75 per cent for passing was, as a standard of work, largely a figment of the imagination. She knows that the difficulty of an examination or test, either as a whole or in respect to any one

question, is to be judged by the success with which students react to it, rather than by the often hurried guess of a single teacher.

The success with which students react to a test or examination question may be thought of as the percentage of the group who can respond satisfactorily. Once this percentage is known, the relative difficulty of each question may be thought of in terms of its σ units of distance from the mean of the total distribution considered, and its difficulty may be expressed in these same units.

For example, a teacher gives a test of 100 questions to 5 classes of 38 students each, or a total of 190. She realizes that there are some differences in difficulty, but is not aware how great these differences are, or which questions are most difficult. She decides to determine the difficulty-value of each question so that she may thereafter arrange them roughly in this order, throwing out those which are too easy and too difficult, and reducing the number of duplicates where questions are of about equal difficulty.

When she has finished her summary for the first five questions, she discovers that:

19	students,	or 10 per cent	answered	question	1	correctly
179	"	about 94.2 per cent	"	"	2	"
152	"	or 80 per cent	"	"	3	"
72	"	about 37.9 per cent	"	"	4	"
114	"	or 60 per cent	"	"	5	"

The computation of the σ difficulty-values of these five questions is shown in Figure 14. Since the mean, or 50 per cent, represents the average difficulty of all the questions

which we might ever ask, and which vary all the way from very easy to very difficult, we shall take it as our starting point or center for the distribution. Thus a question which

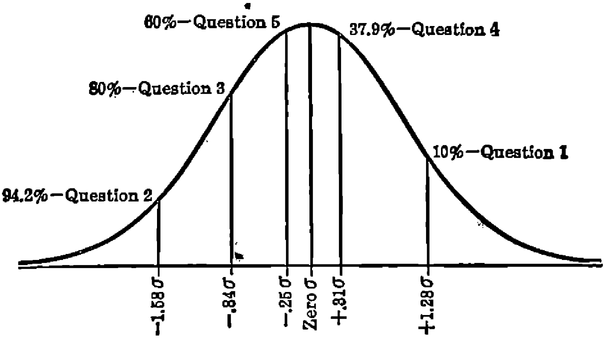


FIG. 14. ILLUSTRATION OF THE COMPUTATION OF DIFFICULTY VALUES FOR FIVE TEST QUESTIONS

was answered correctly by 50 per cent of the class and missed by 50 per cent of the class would have average difficulty, expressed as zero σ difficulty. Question 1, answered by only 10 per cent, is clearly more difficult than this average question, having 40 per cent of the area of the curve between it and the mean. If we look up this area, 40 per cent, in Table XVII, page 115, we find it to correspond to a value of 1.28 σ. Since it is of more than average difficulty we give it a plus sign, and say the question has +1.28 σ units of difficulty. The difficulty of question 4, answered by 37.9 per cent of the class, is found in the same manner, and turns out to have +.31 σ units of difficulty.

With but slight modification, the same process is used in computation on the opposite side of the mean. Instead of

subtracting the percentage value of the questions from 50 in each case, the 50 is subtracted from the percentage of students who answered the question. Thus question 3 was answered correctly by 80 per cent of the 190 students. Therefore 30 per cent of the distribution lies between the mean and the location of this question in the whole distribution. An area which makes 30 per cent of a distribution on one side of the mean is bounded by a perpendicular erected at a distance of .84 σ from the mean, according to Table XVII. The difficulty value of question 3 is, therefore, $-.84\ \sigma$.

> *Problem:* At this point, the student should make the computations in determining the difficulty values of questions 2, 4, and 5.

Giving scale-values to questions. Once the difficulty-value has been determined for each question, it may be placed in its proper order in a so-called scale or measuring device in which the difficulty-value of each task is known, and in which the difference in difficulty between succeeding tasks is as nearly equal as possible.

The σ values of these five questions may now be rearranged in a more convenient form. The mean, as the starting point of the computations, is of course zero σ. Above it, the σ's are positive, and below it negative. If we choose a zero point 3 σ's below the mean (after the difficulty computations have been made), the mean becomes 3 σ. With this again as our starting point, we now have the following scale values for each question:

Question	1	(+ 1.28 σ)	becomes	4.28 σ
"	4	(+ .31 σ)	"	3.31 σ
"	5	(− .25 σ)	"	2.75 σ
"	3	(− .84 σ)	"	2.16 σ
"	2	(− 1.58 σ)	"	1.42 σ

This enables us to count difficulty from a zero point three σ units below the mean, and thus keep all our units in positive terms.

After a teacher understands the process of scaling a test or examination, she should be in a position to rate pupils much more scientifically; and she should have a more objective and scientific attitude toward attempts to evaluate teaching.

The product scale. In objective tests in arithmetic or spelling, answers can usually be graded as right or wrong. A drawing or a piece of English composition, however, offers a new problem. We do not say that it is right or wrong, but we compare it with samples on a scale (if one is available) and assign it a value near that one which it most closely resembles.

Just as problems, words, or other tasks in a test can be scaled (difficulty-value determined), samples of writing or drawing also can be scaled. The problem, however, is somewhat different because here we must rely on the opinion of judges. The process of scaling samples for writing, drawing, or other product scales is as follows: Get as representative a group of samples as possible (samples of writing, for instance, ranging from poorest to best). Have a large number of (presumably) competent judges arrange them in

order of merit. Reduce to per cents the number of times each sample is ranked above each other sample, and reduce these per cents to σ units. These σ units represent differences of quality or merit between samples.

Teachers will not ordinarily scale tests or develop product scales, but a knowledge of how to do so will contribute to a more intelligent utilization of what has been done by others.

Significance of the normal-curve hypothesis. Throughout the discussion of the normal curve, it has been referred to as a hypothesis because we have no proof that a graph of the data of biology, psychology, or education will in every case constitute a normal curve. On the other hand, the evidence from these fields is sufficient to constitute a very strong presumption in favor of this hypothesis, and many scientists have come to accept, with some reservations, the view that distributions of traits and abilities from representative groups tend to be symmetrical, or normal.

When we examine distributions, therefore, any serious departure from a normal curve, whether it be skewed, flattened, or too narrow, is, in general, interpreted that the traits or abilities measured do not represent a random sampling of such traits and abilities among pupils. Consequently, if we wish to be sure that our computations of central tendency and variability are accurate, we must measure these traits or abilities in a sufficient number of additional pupils to obtain a normal distribution. When we have reached this point, the addition of more measurements of the same kind does not affect the sizes of central tendency and variability already computed.

Another use is made of this hypothesis in connection with the classification of pupils. If a test has been properly devised for a group of a given average ability, and it is applied to a normally distributed group of pupils of this average ability, the resulting scores should naturally form a normal distribution. On the other hand, if the resulting distribution is badly skewed, we are inclined to suspect the homogeneity of the group of pupils; i.e., we may feel that they are not sufficiently alike to be taught together, and that a need for reclassification exists in order to take care of those pupils who show exceptional divergence from the rest of the group.

If we start with the knowledge (gained from a number of previous tests) that we have within a certain group of pupils a normal distribution with respect to a certain ability, we may use this knowledge in choosing or improving a new test in this ability. Under these conditions a properly devised test for this group would give a normal distribution of scores or measures. If, however, the resulting scores depart seriously from the normal curve, we have a right to suspect the adequacy of the test. Under these conditions, it may lack discrimination (the power satisfactorily to detect differences in the ability), or it may be too easy or too difficult for the particular group.

Bi-modality. Finally, we may consider the significance of bi-modality. Occasionally, a bi-modal distribution may result from a defective test, but ordinarily it means that two traits or two aspects of a trait instead of one are being measured. In a certain study, the distribution of the

heights of boys appeared to be bi-modal. The distribution of heights of girls in the same study revealed a decided central tendency. A recheck of the data revealed the fact that what was regarded as height data of a larger group of boys really included the height data for both girls and boys. Each tabulated separately yielded rather normal distributions, but combined, they yielded a bi-modal distribution.

TRUE–FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. Biological data tend to follow very closely the normal curve.
2. The normal-curve concept is fundamental in practically all statistical work in education.
3. If a good test gives a very flat distribution it is an indication of very high variability and perhaps poor classification of students.
4. In describing scores in terms of their places in the normal curve, it is safe to consider minus 3 σ as the zero point and plus 3 σ as the maximum possible score.
5. If a distribution is symmetrical it is normal.
6. The percentage of cases corresponding to the various σ unit values will not necessarily be the same in all normal curves.
7. The normal curve is always symmetrical and bell-shaped.
8. The true basis of the normal curve is mathematical, rather than biological.
9. It is impossible to have two normal curves with one quite flat and the other quite tall.
10. The relative difficulty of examination questions may be determined by expressing them in σ units.
11. Normal-curve systems of marking almost invariably divide the base line into five σ units.
12. Each successive σ distance from the mean takes in an equal number of cases.

13. The only true normal-curve system of giving marks involves 7 per cent A, 24 per cent B, 38 per cent C, 24 per cent D, and 7 per cent E or F.
14. In scaling test items for difficulty, the procedure is to locate each one on the basis of its place in the normal curve.
15. The normal curve reaches the base line at three σ units from the mean.
16. The fact that the normal curve is derived mathematically constitutes good proof that teachers' marks should be so distributed.
17. The fact that the normal curve is based on chance indicates that any use made of it is likely to be quite uncertain.
18. A bi-modal distribution always shows that the testing instrument is inadequate.
19. The greatest advantage of distributing marks according to the normal curve is that all instructors mark uniformly.
20. There is a constant ratio in the curve of a normal distribution, between any given σ unit and the per cent of the distribution cut off by a perpendicular erected at that distance from the mean.

QUESTIONS AND EXERCISES

1. What evidence is there that frequency distributions of measurements of ability follow the normal curve?
2. What is the source of the theoretically correct curve?
3. Give the rule for expanding a binomial.
4. What is the relation of coin-tossing experiments to the normal-curve hypothesis?
5. If a polygon is plotted from normally distributed data, describe the relationship which exists between the areas enclosed by perpendiculars erected on the base line and the distances of each such perpendicular from the mean.
6. Describe the significance of the normal-curve hypothesis as it applies to education.

7. Expand $(x + y)^7$, and plot the result. Do it by rule and also by ordinary algebraic multiplication. Explain why each yields the same result.

8. A director of research obtained the scores of 10,000 pupils on an arithmetic test. The mean was 36 and the σ was 8.
   a. About how many pupils probably had scores below 28?
   b. How many pupils probably had scores above 44?
   c. How many pupils probably had scores between 30 and 40?

9. A random selection of 1000 samples of writing was scored on the Ayres scale, yielding a mean of 55 and a σ of 10.
   a. About how many students probably scored better than 80?
   b. About how many were probably found between 60 and 80?
   c. About how many probably scored below 50?

10. A junior high school teacher gave an informal test to 370 pupils in geography. How many marks of "Fail" should be given if the passing point is set at $-1.5\ \sigma$? How many if it were at $-2.5\ \sigma$?

11. In a test administered to 650 students in 9A English, questions A, D, L, and T were answered satisfactorily by 65, 125, 400, and 575 students respectively. How does the difference in difficulty between questions A and D compare with the difference in difficulty between questions L and T?

12. A group of 22 judges, in attempting to develop a writing scale, rank many samples of writing. Sample X is ranked better than sample Y by 18 of them; sample Y is ranked better than sample Z by 15 of them. What is the difference in merit between X and Y, and Y and Z in terms of σ units?

CHAPTER IX

RELIABILITY

Statistical reliability. If we find that 25 boys twelve years of age have a mean weight of 81 pounds, are we warranted in saying that 81 is the normal weight for twelve-year-old boys? In other words, are we at liberty to generalize from an observed group of cases? If so, how can we be sure that the generalizations drawn will be applicable to the thousands of cases that were not observed or measured? This is the problem we confront when we undertake the study of reliability.

The reliability of a statistical quantity tells us the extent to which we can accept it as being truly representative of an infinite number of cases. In other words, reliability, in the sense in which it is used in this chapter, means the extent to which we can predict from a limited observed sampling the outcome that would be expected if we were to examine and observe an infinitely large number of cases, or if we should observe all the cases there are in the world. Thus, going back to our illustration, the reliability of our mean of 81 pounds for twelve-year-old boys would tell us how safe we would be in publishing 81 as the norm for twelve-year-old boys throughout the United States, or throughout the world.

What reliability does not include. The beginner is likely to confuse reliability with a number of other terms, such as

accuracy, validity, and representativeness of sampling. Let us, therefore, clear up the meaning of our terms before we go further.

Reliability does not refer to accuracy of measurement, nor to accuracy of mathematical computations. Errors in such things as these may be numerous and great and may spoil the value of a study, but they are not cases of unreliability in the strict sense.

Nor is reliability the same as validity. If an investigator gives a test to measure intelligence but really measures linguistic training, he has sacrificed validity, but his invalid results might be highly reliable, because if he were to repeat his study for additional cases he would get similar results.

Reliability is not the same as authenticity. Data secured from experimental studies may be very authentic, in that the investigator has personally observed the cases and carefully supervised the study, yet the results may be very unreliable, because he included only a few cases in the experiment! Data collected by questionnaire from 4000 persons might be very doubtful on grounds of truthfulness of the reports, and yet be very reliable, in the statistical sense, because of the large number of cases included in the study.

Finally, reliability is not the same as representativeness. A study based on 10,000 school children in the fourth grade ordinarily would be considered highly reliable, because of the large number of cases; but if the 10,000 cases were all selected from one-room rural schools they could by no means be considered as representative of American fourth-

grade children. In other words, we have high reliability but lack representativeness. If, however, we are studying rural schools only, then we could say they were representative of rural schools, provided other conditions were typical.

Reliability an application of the theory of chance. Reliability is simply an aspect of the laws of probability, and is dependent upon the number and variability of the cases observed. For example, five coins were tossed, and yielded three heads. They were tossed again and yielded two heads. From these two cases of coin tossing we might generalize that on the average, five coins would yield 2.5 heads. This generalization is questionable because there were only two samplings, but it is somewhat strengthened by the fact that we got three heads one time and two the other instead of four one time and one the other. In other words, the generalization is safest (most reliable) when the cases are numerous and have low variability.

But we need not stop with two tosses. We could generalize more safely from 20 tosses, which yielded an average of 2.6 heads. Forty tosses gave an average of 2.52; 60 tosses an average of 2.47; 80 tosses an average of 2.56; and 100 tosses an average of 2.55.

We could go on tossing the coins for 10,000 times or more, and thus be very certain of the behavior of the coins, but it would be more economical to make a mathematical computation on the basis of our 100 tosses and thus save the labor of the actual coin handling. In other words, the mathematical computation of reliability is a short method

for estimating the probability that a prediction based on a few observed cases will prove true when verified by observation of an infinite number of cases. That prediction is possible because of the very definite laws of chance or probability.

Reliability dependent upon variability. Reliability is inversely proportional to variability. Thus, referring again to our mean of 81 pounds for twelve-year-old boys' weights, clearly it would be safer to predict weights of boys in general if the σ of the distribution of weights were 10 pounds than if it were 40 pounds, because this fact would signify a much more homogeneous grouping of the weights. If the observed cases are close together in the scale, this indicates that the unobserved cases will probably be concentrated within a small range also. If, however, the observed cases are scattered over a very wide range, there is less certainty as to where the average of the unobserved cases will be. Thus it is clear that reliability increases as variability decreases, and vice versa.

Reliability dependent upon number of observed cases. The other factor determining reliability is the number of cases observed. Reliability increases as the number of cases increases. Thus the mean weight of 50 boys is a more reliable indication of the true weight of boys in general than is the mean weight of 25 boys, but not nearly so reliable, in general, as that of 500 boys. As a rule, the greater the number of cases we observe, the more certain we are that additional observations will not reveal a radically different result from what we have already secured.

Standard error as a unit for the measurement of reliability. Reliability is ordinarily measured either in terms of standard error (to be represented hereafter in this book by the Greek letter epsilon, ϵ), or in terms of probable error (usually written as *P.E.*). Standard error is numerically the same as standard deviation, but carries a different interpretation. ϵ stands for probability, prophecy, or prediction, whereas σ stands for observation, history, or fact. ϵ is predictive, whereas σ is descriptive. σ is used as a measure of variability, or dispersion; whereas ϵ is used in generalizing from observed cases or data to the entire population of which the observed cases were a representative sampling. This will be more apparent as we proceed in the chapter.

Reliability of a mean. Let us now take up the method of determining the reliability of the mean. In other words, let us learn how to find and interpret the standard error of the mean, ϵ_M.

We shall illustrate this concept first by reference to the experiment with the coins. The first time we tossed five coins twenty times, the mean number of heads which came up was 2.60; a second set of 20 tosses yielded a mean of 2.45 heads; a third set, a mean of 2.35 heads; a fourth set, 2.85 heads; and a fifth set, 2.50 heads. Notice that each of these means was computed on a single set of 20 tosses and that they differ somewhat from the means previously obtained, which were computed from 20, 40, 60, 80, and 100 tosses, respectively. As a result of our five samplings, we have these five actual means. It is apparent that if we made an in-

finite number of additional sets of tosses, we should get a large number of additional means, all of them grouped more or less closely around 2.50. We could make a frequency table for these means, find their standard deviation, and thus determine the limits within which two thirds of them actually fell. This would be the *standard deviation of the actual means*, or the σ_M. It would be shorter and easier, however, to arrive at approximately the same result mathematically by computing the *standard error of the mean*, ϵ_M, based on our first 20, 60, or 100 cases, in a manner which we shall presently describe.

The number of heads up in five coins varied from none to five, the largest numbers being two and three. The mean for 100 tosses was 2.55, and the σ 1.09. From N and σ we compute ϵ_M. Reliability is increased by a low variability and by a large number of cases; the formula takes account of these two factors by putting σ in the numerator and $\sqrt{N}$ in the denominator.

Formula:[1]

$$\epsilon_M = \frac{\sigma}{\sqrt{N}}$$

In the present problem, therefore, we substitute our actual numbers in the formula and obtain:

$$\epsilon_M = \frac{1.09}{\sqrt{100}}$$
$$= .11$$

What does this ϵ_M of .11 tell us? It indicates that if we should toss five coins indefinitely, getting the mean number of heads for each 100 tosses, approximately two thirds

[1] For derivation, see Kelley, T. L., *Statistical Method*, pp. 82–83.

(68.26 per cent) of all the means so found would be within .11 of the observed mean, 2.55. In other words, the chances are two to one that the true mean lies within .11 units of 2.55, or between 2.44 and 2.66. The ϵ_M therefore establishes limits within which we can predict the true mean for all cases on the basis of a few observed cases.

A second illustration of the standard error of the mean. Let us take a second illustration in order to help clarify what we have done so far. A class of 50 sixth-grade pupils obtained a mean score of 65 with a σ of 10 on a test in arithmetic. The teacher wished to know the most probable limits within which the average of all sixth-graders would fall if the test were applied throughout the school. Substituting in the formula she obtained:

$$\epsilon_M = \frac{10}{\sqrt{50}}$$
$$= 1.41$$

She was able to say, therefore, that in so far as her class of 50 was representative of the school as a whole, the chances were approximately 2 to 1 that the school average for sixth-graders would lie between $65 + 1.41$ and $65 - 1.41$, or between 66.41 and 63.59.

Since a distance of 3 ϵ_M each way from the mean takes in over 99 per cent of all the cases in a normal distribution, if we take 3×1.41 on each side of the mean of 65, we will set limits within which we are *practically certain* to find the true mean of an infinite number of samplings. Thus, in this case, we are practically certain that the true mean will lie between 60.77 and 69.23.

Determining the chances that the true mean is above a given point. From what we learned, it is readily seen that we can determine the chances that the true mean is above any given point. Thus, in the previous example, if 62 is the national sixth-grade norm on that particular test, we can find how many ϵ_M units 62 is below our observed mean of 65. Since ϵ_M is 1.41, a score of 62 is $3/1.41$ ϵ_M units, or 2.1 ϵ_M units below the observed mean. By referring to Table XIX we find that the chances are more than 98 out of 100 (column 2) that a case would not be more than 2.1 ϵ_M units below the observed mean. In other words, the chances are 54.8 to 1 (column 4) that there is a real superiority in this school over the national norm, and that the true mean of sixth-grade achievement in the school is above the national norm of 62. (This, of course, is based on the assumption that the pupils tested are typical of all the sixth-graders in the school.)

The number of cases necessary to produce a given degree of reliability. The reliability formula is useful not only in computing ϵ_M but in determining how many measurements must be made to insure that an obtained mean does not depart from the true mean by more than some predetermined amount.

For example, in attempting to determine the ability of high-school seniors in handwriting, 25 papers were scored (Ayres Scale, Gettysburg Edition), and the mean score was found to be 55 with a σ of 20. The ϵ_M was $20/\sqrt{25}$, or 4. This information indicated that on the basis of 25 measurements, 68 per cent of the future means computed on dis-

Table XIX. Chances that Measures will not Fall Below Various Standard Error Unit Values

(The table has been adapted from McCall's *How to Experiment in Education*, p. 101.)

Standard error units	Per cent above	Per cent below	Ratio above to below[1]	Standard error units	Per cent above	Per cent below	Ratio above to below[1]
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
.00	50.00	50.00	1.0	2.55	99.46	.54	184
.05	51.99	48.01	1.1	2.60	99.53	.47	212
.10	53.98	46.02	1.2	2.65	99.60	.40	249
.15	55.96	44.04	1.3	2.70	99.65	.35	285
.20	57.93	42.07	1.4	2.75	99.70	.30	332
.25	59.87	40.13	1.5	2.80	99.74	.26	384
.30	61.79	38.21	1.6	2.85	99.78	.22	453
.35	63.68	36.32	1.8	2.90	99.81	.19	525
.40	65.54	34.46	1.9	2.95	99.84	.16	625
.45	67.36	32.64	2.1	3.00	99.865	.13	770
.50	69.15	30.85	2.2	3.05	99.886	.11	900
.55	70.88	29.12	2.4	3.10	99.903	.097	1000
.60	72.57	27.43	2.6	3.15	99.918	.082	1200
.65	74.22	25.78	2.8	3.20	99.931	.069	1450
.70	75.80	24.20	3.1	3.25	99.942	.058	1700
.75	77.34	22.66	3.4	3.30	99.952	.048	2100
.80	78.81	21.19	3.7	3.35	99.960	.040	2500
.85	80.23	19.77	4.1	3.40	99.966	.034	2900
.90	81.59	18.41	4.4	3.45	99.972	.028	3600
.95	82.89	17.11	4.8	3.50	99.977	.023	4300
1.00	84.13	15.87	5.3	3.55	99.981	.019	5400
1.05	85.31	14.69	5.8	3.60	99.984	.016	6200
1.10	86.43	13.57	6.4	3.65	99.987	.013	7700
1.15	87.49	12.51	7.0	3.70	99.989	.011	9100
1.20	88.49	11.51	7.7	3.75	99.9912	.009	11000
1.25	89.44	10.56	8.5	3.80	99.9928	.007	14000
1.30	90.32	9.68	9.3	3.85	99.9941	.0059	17000
1.35	91.15	8.85	10.3	3.90	99.9952	.0048	21000
1.40	91.92	8.08	11.4	3.95	99.9961	.0039	26000
1.45	92.65	7.35	12.6	4.00	99.9968	.0032	31000
1.50	93.32	6.68	14.0	4.05	99.9974	.0026	38000
1.55	93.94	6.06	15.5	4.10	99.9979	.0021	48000
1.60	94.52	5.48	17.2	4.15	99.9983	.0017	59000
1.65	95.05	4.95	19.2	4.20	99.9987	.0013	77000
1.70	95.54	4.46	21.5	4.25	99.9989	.0011	91000
1.75	95.99	4.01	24.0	4.30	99.99915	.0009	110000
1.80	96.41	3.59	26.8	4.35	99.99932	.0007	140000
1.85	96.78	3.22	30.1	4.40	99.99946	.0005	200000
1.90	97.13	2.87	33.8	4.45	99.99957	.00043	230000
1.95	97.44	2.56	38.0	4.50	99.99966	.00034	290000
2.00	97.72	2.28	42.8	4.55	99.99973	.00027	370000
2.05	97.98	2.02	48.5	4.60	99.99979	.00021	480000
2.10	98.21	1.79	54.8	4.65	99.99983	.00017	590000
2.15	98.42	1.58	62.2	4.70	99.99987	.00013	770000
2.20	98.61	1.39	70.9	4.75	99.99990	.00010	1000000
2.25	98.78	1.22	81.0	4.80	99.99992	.00008	1200000
2.30	98.93	1.07	92.4	4.85	99.999938	.000062	1600000
2.35	99.06	.94	105.0	4.90	99.999952	.000048	2100000
2.40	99.18	.82	121.0	4.95	99.999963	.000037	2700000
2.45	99.29	.71	140.0	5.00	99.999971	.000029	3400000
2.50	99.38	.62	160.0				

[1] This column represents the number of chances to one that a case would be above the corresponding ε value. Thus, for 1.50 ε units the chances are 14 to 1, and for 5 ε the chances are 3,400,000 to 1.

tributions of twelfth-grade writing scores would fall between the limits 51–59. If we decide that this mean is not sufficiently reliable, and we wish to determine how many samples of the handwriting of high-school seniors must be included before the mean which is obtained (whatever it is) will not deviate (68 per cent of the time) from the true mean of the handwriting of all seniors by more than 1 point, the computation is made by substituting the desired ϵ_M in the formula together with the σ and solving for the item, $\sqrt{N}$. The computation becomes $1 = 20/\sqrt{?}$ and the N required is 400. This information indicates that a mean obtained from 400 measurements (instead of 25) will, 68 per cent of the time, deviate not more than 1 point from the true mean of handwriting ability of high-school seniors. This would be true, of course, only in so far as these 400 are representative of high-school seniors in general.

Reliability of a difference between two means. Sometimes we wish to compare two methods of teaching, or two other similar items, and decide which one to recommend to teachers at large. Thus an investigator conducted a study to compare the reasoning ability of freshman boys and freshman girls. The problem was to discover whether boys in general reason better than girls, as indicated by the scores of 290 boys and the 179 girls whose scores were available. The scores for the two groups were as follows: The 290 boys averaged 9.81 with a σ of 3.74; and the girls averaged 8.62 with a σ of 3.50. The difference between the two means, $9.81 - 8.62$, was 1.19. Is this difference large enough so that it could not be due to chance, and may we interpret

it as indicating that boys in general reason better than girls? Computing its reliability will give the answer.

There is no question about the *observed* difference. We know that it is 1.19. We are not sure of the size of the *true* difference, however, or the one that we would get if we were to test an infinite number of boys and of girls. In reality we are subtracting 8.62 more-or-less from 9.81 more-or-less, and we find the difference to be 1.19 more-or-less. But how much is the "more-or-less"? To answer this we get the ϵ_M for each mean, by the formula already given, finding ϵ_{M_1} (for boys) to be $3.74/\sqrt{290}$, or .22, and ϵ_{M_2} (for girls) to be $3.5/\sqrt{179}$, or .26. How much, then, is the difference between 9.81, ϵ_M .22, and 8.62, ϵ_M .26? Since in an infinite number of comparisons the differences would distribute themselves approximately according to a normal curve, we may answer this question by computing the standard error of the difference. The formula is:

$$\epsilon_D{}^* = \sqrt{(\epsilon_{M_1})^2 + (\epsilon_{M_2})^2}$$

Substituting in the formula, we have, for the present problem:

$$\begin{aligned}\epsilon_D &= \sqrt{(.22)^2 + (.26)^2}\\ &= .34\end{aligned}$$

Having found ϵ_D to be .34, we now know that the chances are 2 to 1 (68.26 per cent) that the true difference will lie within .34 units from the observed difference of 1.19. In other words, in so far as our test is an indication of the

* For a brief practical treatment of this, see McCall, W. A., *How to Experiment in Education*, pp. 150–154.

reasoning ability of boys and girls, we are able to say that the chances are 2 to 1 that freshman boys will exceed girls on this test by an amount between .85 and 1.53.

Determining the chances that the true difference is above zero. The case just described still leaves us a bit uncertain, because we want to know what the chances are that boys will always have a difference in their favor. If the chances are 2 to 1 that the true difference lies between .85 and 1.53, what are the chances that it lies somewhere above zero? We can answer this question by finding the ratio between the difference and its standard error. The formula is as follows:

$$\text{Ratio} = D/\epsilon_D$$

This ratio tells us how many ϵ_D units the observed D is above zero. Thus, in the above example the observed D is 1.19, and its ϵ_D is .34. Substituting in the formula, we have:

$$\text{Ratio} = 1.19/.34, \text{ or } 3.5$$

This tells us that the observed difference is 3.5 ϵ_D units away from zero, or that the zero point is 3.5 ϵ_D units below the observed difference. By referring to Table XIX, column 4, we find that the chances are about 4300 to 1 that the true difference would not be 3.5 ϵ_D units below the observed difference, and we are therefore justified in saying that, in so far as our tests are valid and our groups representative, we are very certain that freshman boys reason better than freshman girls.

This may be further clarified by means of Figure 15, in which we have drawn a curve and located the various ϵ_D

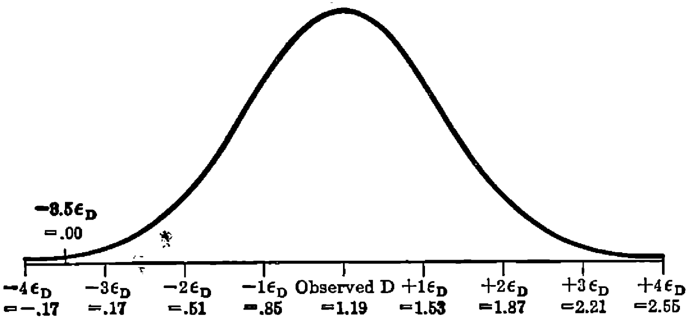


FIG. 15. ILLUSTRATION OF THE BASIS FOR CONVERTING AN OBSERVED DIFFERENCE INTO A STATEMENT OF CHANCES THAT THE TRUE DIFFERENCE IS ABOVE ZERO

The ratio D/ϵ_D tells how many ϵ_D units above zero the observed D is, and this number can then be converted into chances, using Table XIX. In the present case the D, 1.19, is 3.5 ϵ_D units above zero, and the chances are about 4350 to 1 that there is a true D greater than zero.

units. By noting where the zero point falls in this curve it is clear that the chances are very high that the true difference would not fall below zero, since the curve has almost reached the base line before we get as far down as zero.

Reliability of the sum of two or more means. The same type of formula that is used to get the reliability of a difference is also good to get the reliability of a sum of two or more means. The formula is:

$$\epsilon_S = \sqrt{(\epsilon_{M_1})^2 + (\epsilon_{M_2})^2}$$

The use of this formula will be illustrated later.

Reliability of the difference between the sums of two sets of means. After adding two pairs of means and getting the reliability of the respective sums, the difference between the sums may be found and its reliability then computed, using this formula:

$$\epsilon_D = \sqrt{(\epsilon_{S_1})^2 + (\epsilon_{S_2})^2}$$

The use of this formula will be illustrated in the next paragraph.

Reliability illustrated as it is involved in a typical rotation experiment. Let us illustrate reliability as it was involved in an actual experimental study made by one of the writers. The problem was to discover whether note-taking during college lectures improved quiz grades. There were two groups, A and B, one of which took notes one day while the other listened; and on the second day they rotated methods so that those who took notes the first day listened the second day, and vice versa. The essential data for the whole experiment are presented in Table XX, together with the necessary computations down to finding the statement of chances. There are no new principles here, but most of the familiar principles are involved in one form or another. This example, if thoroughly understood, will prove valuable as a model, and will serve to organize the work in cases where the statistical procedure becomes complex. Models for various other experimental situations are found in McCall's *How to Experiment in Education.*

Reliability of other statistical quantities. Before leaving this topic we should state that the reliability of numerous other statistical quantities can be computed, and that, in general, the interpretation of the results of such computations is similar to the interpretation of those already explained. Thus it is possible to determine the reliability of a median by the formula:

$$\epsilon_{Md} = 5\,\sigma/4\sqrt{N}$$

TABLE XX. ILLUSTRATION OF THE COMPUTATION OF VARIOUS QUANTITIES AS THEY ARE INVOLVED IN A ROTATION EXPERIMENT

(The data are scores of college students in an experiment comparing note-taking and listening, as measured in terms of later quiz grades.)

Items	Group A		Group B	
	First day notes	Second day listening	First day listening	Second day notes
	(1)	(2)	(3)	(4)
N	63	63	48	48
M	8.58	6.41	6.21	6.75
σ	2.39	2.69	2.47	2.68
ϵ_M	.301	.348	.356	.386

S_1, Notes $= M_1 + M_4 = 8.58 + 6.75 = 15.33$
S_2, Listening $= M_2 + M_3 = 6.41 + 6.21 = 12.62$
$D = S_1 - S_2 = 15.33 - 12.62 = 2.71$

$$\epsilon_{S_1} = \sqrt{(\epsilon_{M_1})^2 + (\epsilon_{M_4})^2}$$
$$= \sqrt{(.301)^2 + (.386)^2}$$
$$= .49$$
$$\epsilon_{S_2} = \sqrt{(\epsilon_{M_2})^2 + (\epsilon_{M_3})^2}$$
$$= \sqrt{(.348)^2 + (.356)^2}$$
$$= .50$$
$$\epsilon_D = \sqrt{(\epsilon_{S_1})^2 + (\epsilon_{S_2})^2}$$
$$= \sqrt{(.49)^2 + (.50)^2}$$
$$= .70$$

Ratio $= D/\epsilon_D = 2.71/.70 = 3.9$

(Chances that true D is above zero, by Table XIX, about 21,000 to 1. Note that the ratio 3.9 means that the observed D is 3.9 ϵ_D units above zero.)

Likewise the reliability of σ may be computed by means of the formula:

$$\epsilon_\sigma = \sigma/\sqrt{2\,N}$$

The reliability of a correlation coefficient, as well as that

of the difference between two correlation coefficients, will be treated in a later chapter.

Probable error as a unit for the measurement of reliability. Just as we can measure distances in miles and also in kilometers, we can measure reliability in standard error or probable error units. The only difference is that *P.E.* is .67449 ϵ. This ratio is constant, just as the ratio of feet to yards is constant at .33333. *P.E.* is, in reality, that fraction of ϵ which represents a 50–50 chance. In other words, whereas ϵ tells us the limits in which two thirds of the chances lie, *P.E.* tells us the limits within which one half of them lie. There is one chance in four that the true measure may be above +1 *P.E.*; one in four that it may fall below −1 *P.E.*, and two in four that it will fall between +1 *P.E.* and −1 *P.E.*

Relative merits of probable error and standard error as units for the measurement of reliability. Which of these units is preferable is a matter of individual taste. Either one will serve the purpose well. The writers advocate, however, that the student of statistics learn them both well enough to understand them when he meets them in educational literature, and learn one of them well enough to use it skillfully and economically in making calculations.

Advantages of the standard error. The writers prefer to use ϵ instead of *P.E.* and have presented it first in this chapter, for the following reasons:

1. *P.E.* is computed by first getting ϵ and then multiplying by the constant, .67449. *P.E.* therefore requires one more step, and introduces one more chance for arithmetical errors to creep into the work.

2. They feel that defining the limits of a statistical quantity in terms of the 2-to-1 chances is more satisfactory than defining them in terms of equal or 1-to-1 chances. This is because the 2-to-1 division point, or one ϵ distance from the mean, is the point where the normal curve changes from convex to concave. The cases which fall between the mean and one ϵ are much more homogeneous than those that fall beyond one ϵ. Describing a distribution in terms of its mean and ϵ_M involves an accounting for all the cases that are most nearly homogeneous, therefore, and ignores the 16 per cent of extremely variant cases at each end of the curve. The use of the 1-to-1 chances, or the *P.E.*, on the contrary, draws a dividing line closer to the mean and takes in some of the rather homogeneous central group of cases and excludes the remainder, along with the highly variant cases that fall farther out than the point of inflection of the curve.

Advantages of the probable error. Opposed to the considerations just presented are a number of significant arguments in favor of using the *P.E.* for measuring reliability:

1. *P.E.* is very widely used in the other sciences, such as astronomy, biology, and economics. There is an advantage in using it in education in order that all research workers will be able to talk the same language.

2. *P.E.* has been used for a longer time, and is more firmly established. We have a larger "investment" in it than in the standard error. Any change of practice now may cause inconvenience and confusion, just as we should have trouble

if the dairies began all of a sudden to sell milk by the liter instead of the quart.

3. *P.E.* is used more extensively in educational literature than ϵ. Consequently, a thorough acquaintance with it is necessary for any one who wishes to read intelligently the writings of those who still prefer to use the *P.E.*

4. We have a convenient and compact notation for the *P.E.* which is lacking for ϵ. Scientific workers are generally agreed that the expression, "Mean $= 56 \pm 2$," indicates that the probable error of the mean is 2. There is no such compact and universally understood symbolism to indicate the ϵ of a measure. It is necessary, therefore, to use some such arrangement as "Mean $= 56$, $\epsilon_M = 3$," or else, in using the form 56 ± 3, to state very definitely that the standard error is meant rather than the probable error.

Formulas for reliability, using P.E. instead of ϵ.[1] All the formulas that were given earlier in the chapter and based on the use of ϵ can be rewritten based on *P.E.* We need only to multiply the appropriate values by the constant, .67449. The formula for the probable error of the mean thus becomes:

$$P.E._M = .67449\epsilon_M$$

We may express the same formula in terms of the original quantities from which ϵ_M is computed, obtaining:

$$P.E._M = \frac{.67449\ \sigma}{\sqrt{N}}$$

In order to determine the chances that the true mean is

[1] For more detailed treatment of these and other *P.E.* formulas, see Holzinger, K. J., *Statistical Methods for Students in Education*, chap. XIII.

greater than zero we find the ratio between the mean and its $P.E._M$ and refer to Table XXI. Thus:

$$\text{Ratio} = M/P.E._M$$

The formula for the probable error of a difference is very similar to that for the standard error of a difference, being:

$$P.E._D = \sqrt{(P.E._{M_1})^2 + (P.E._{M_2})^2}$$

To get the chances that the true difference is greater than zero, we get the ratio between the difference and its $P.E._D$ and refer to Table XXI. Thus:

$$\text{Ratio} = D/P.E._D$$

The probable error of a sum of two means is derived from a similar formula:

$$P.E._S = \sqrt{(P.E._{M_1})^2 + (P.E._{M_2})^2}$$

The probable error of a median is found by the formula:

$$P.E._{Md} = 1.25\ P.E._M$$

The formula for the probable error of a standard deviation is:

$$P.E._\sigma = \frac{.67449\sigma}{\sqrt{2N}}$$

The probable error of a correlation coefficient will be discussed in Chapter XI.

The use of the P.E. table to determine chances. We have learned how to use Table XIX in determining chances that measures will not fall below various ϵ unit values. It now remains for us to learn how to use a similar table prepared on the basis of *P.E.* instead of ϵ units. Table XXI will serve this purpose.

Table XXI. Chances that Measures will not Fall below Various *P.E.* Unit Values

(The table has been adapted from Rugg's *Statistical Methods Applied to Education,* p. 391.)

P.E. Units	Per cent above	Per cent below	Ratio above to below [1]	*P.E.* Units	Per cent above	Per cent below	Ratio above to below [1]
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
.00	50.00	50.00	1.0	1.70	87.42	12.58	6.9
.05	51.35	48.65	1.1	1.75	88.11	11.89	7.4
.10	52.69	47.31	1.1	1.80	88.96	11.04	8.1
.15	54.03	45.97	1.2	1.85	89.39	10.61	8.4
.20	55.36	44.64	1.2	1.90	90.00	10.00	9.0
.25	56.70	43.30	1.3	1.95	90.57	9.43	9.6
.30	58.02	41.98	1.3	2.00	91.13	8.87	10.3
.35	59.33	40.67	1.5	2.05	91.66	8.34	11.0
.40	60.63	39.37	1.5	2.10	92.17	7.83	11.8
.45	61.93	38.07	1.6	2.15	92.65	7.35	12.6
.50	63.21	36.79	1.7	2.20	93.11	6.89	13.5
.55	64.47	35.53	1.8	2.25	93.54	6.46	15.4
.60	65.71	34.29	1.9	2.30	93.96	6.04	15.5
.65	66.95	33.05	2.0	2.35	94.35	5.65	16.7
.70	68.16	31.84	2.1	2.40	94.72	5.28	17.9
.75	69.35	30.65	2.3	2.45	95.08	4.92	19.3
.80	70.53	29.47	2.4	2.50	95.41	4.59	20.8
.85	71.68	28.32	2.5	2.55	95.73	4.27	22.4
.90	72.91	27.09	2.7	2.60	96.02	3.98	24.1
.95	73.92	26.08	2.8	2.65	96.31	3.69	26.1
1.00	75.00	25.00	3.0	2.70	96.57	3.43	28.2
1.05	76.06	23.94	3.2	2.75	96.82	3.18	30.4
1.10	77.09	22.91	3.4	2.80	97.05	2.95	32.9
1.15	78.10	21.90	3.6	2.85	97.27	2.73	35.6
1.20	79.08	20.92	3.8	2.90	97.48	2.52	38.7
1.25	80.04	19.96	4.0	2.95	97.67	2.33	42.0
1.30	80.97	19.03	4.3	3.00	97.85	2.15	45
1.35	81.88	18.12	4.5	3.05	98.02	1.98	50
1.40	82.75	17.25	4.8	3.10	98.17	1.83	54
1.45	83.60	16.40	5.1	3.15	98.31	1.69	58
1.50	84.41	15.59	5.4	3.20	98.45	1.55	63
1.55	85.21	14.79	5.8	3.25	98.58	1.42	69
1.60	85.97	14.03	6.1	3.30	98.70	1.30	76
1.65	86.71	13.29	6.5	3.35	98.81	1.19	83

[1] This column represents the number of chances to one that a measure would be above the corresponding *P.E.* value. Thus for 1.00 *P.E.* the chances are 3 to 1; and for 5.95 *P.E.* they are 33,000 to 1.

TABLE XXI (*continued*)

P.E. Units	Per cent above	Per cent below	Ratio above to below	*P.E.* Units	Per cent above	Per cent below	Ratio above to below
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
3.40	98.91	1.09	91	4.70	99.92	.08	1250
3.45	99.00	1.00	99	4.75	99.93	.07	1400
3.50	99.09	.91	109	4.80	99.94	.06	1700
3.55	99.17	.83	119	4.85	99.946	.054	1850
3.60	99.24	.76	130	4.90	99.952	.048	2100
3.65	99.31	.69	144	4.95	99.957	.043	2300
3.70	99.37	.63	158	5.00	99.962	.038	2600
3.75	99.43	.57	175	5.05	99.966	.034	2900
3.80	99.48	.52	191	5.10	99.971	.029	3400
3.85	99.53	.47	212	5.15	99.974	.026	3800
3.90	99.57	.43	232	5.20	99.977	.023	4300
3.95	99.61	.39	255	5.25	99.980	.020	5000
4.00	99.65	.35	285	5.30	99.982	.018	5600
4.05	99.68	.32	311	5.35	99.984	.016	6200
4.10	99.71	.29	344	5.40	99.986	.014	7100
4.15	99.74	.26	384	5.45	99.988	.012	8300
4.20	99.77	.23	434	5.50	99.990	.010	10000
4.25	99.79	.21	475	5.55	99.991	.009	11000
4.30	99.81	.19	525	5.60	99.992	.008	12500
4.35	99.83	.17	588	5.65	99.993	.007	14000
4.40	99.85	.15	665	5.70	99.994	.006	17000
4.45	99.87	.13	770	5.75	99.995	.005	20000
4.50	99.88	.12	830	5.80	99.9955	.0045	22000
4.55	99.89	.11	910	5.85	99.9960	.0040	25000
4.60	99.90	.10	1000	5.90	99.9965	.0035	29000
4.65	99.91	.09	1100	5.95	99.9970	.0030	33000

In using Table XXI, the essential first step is to get the ratio of the statistical measure to its *P.E.* Thus, if a mean is 24 and its *P.E.* is 6, the mean is 24/6, or 4 *P.E.* units above zero. We find 4.00 *P.E.* in the table and note in the last column that the chances are 285 to 1 that the true mean will be above zero. Or, if the difference between two means is 10 and the probable error of the difference ($P.E._D$) is 5,

the difference is 10/5 or 2 *P.E.* units above zero. By consulting Table XXI we find that the chances are 10.3 to 1 that the true difference is greater than zero.

TRUE-FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. If data did not conform to the normal curve our methods of determining reliability would have to be completely changed.
2. *P.E.* is not as accurate as standard error.
3. If the directions for a test were improperly given this would decidedly reduce the reliability of the test scores.
4. A measure has the greatest reliability when its standard error is large.
5. The reliability of the measure of a trait is the extent to which it approaches the true measure of that trait.
6. Experimental data have higher reliability than questionnaire data.
7. Number of cases is a less important factor in reliability than variability of the scores.
8. Reliability formulas are to detect sources of error introduced by teacher and pupil variation, health, attendance, etc.
9. As N goes up, reliability goes down.
10. The formula for the standard error of the mean is a short method of finding what the mean would be if we observed innumerable cases.
11. It would be impossible to measure reliability in terms of quartile deviation.
12. Standard error tells us how much faith we can put in computations based on a small number of cases.
13. Standard error and standard deviation are numerically the same.
14. The standard error of the mean equals $\sigma/\sqrt{N}$.

15. The standard error of a difference is found by adding the square roots of the differences.
16. Standard error is always found from probable error.
17. The number of cases needed to produce a given reliability can be determined by formula.
18. To find the chances that the true difference is greater than zero, you divide the difference by its probable or standard error and refer to the appropriate table.
19. Reliability is inversely proportional to variability.
20. Probable error is one half the size of standard error.

QUESTIONS AND EXERCISES

1. What is meant by the reliability of a measure?
2. What particular type of error is measured by reliability formulas?
3. Name several types of errors in measurements not measured by reliability formulas.
4. If the standard error of a mean of 15 is 1.5, what does this information tell us about the total range of means which we might expect if a large number of additional means were computed from similar data? What are the chances that the true M is above 12? What would the chances be that the true mean is above 12 if $P.E._M$ were 1.5?
5. Give the formulas for computing ϵ_M and $P.E._M$.
6. In what manner does the reliability of a mean vary in its relation to the σ and N?
7. Explain what is meant by the reliability of the difference between two means.
8. Give the formulas for ϵ_D and $P.E._D$ and explain in what respect they are similar to each other.
9. The ϵ_D for an obtained difference of 4 is 1.25. What does this information tell us of the total range of differences we might expect if a large number of additional pairs of means were computed and their obtained differences tabulated?

What are the chances that the true D is above zero? What would the chances be that the true difference is above zero if D were 4 and $P.E._D$ were 1.25?

10. An arithmetic test given to 100 sixth-grade pupils yielded a mean of 25.8, and a σ of 3.1. If this test is given to an infinitely large number of groups of sixth-grade pupils:
   a. What per cent of the time will the mean be 26.5 or greater?
   b. What per cent of the time will the mean be less than 25?
   c. Between what limits will it fall approximately 68% of the time?

   (Check your work by using ϵ and $P.E.$ and see if you get the same results.)

11. An eighth-grade class earned a mean of 87 on a test on which a ninth grade earned 96. There were 100 eighth-graders and 81 ninth-graders. The σ for eighth-graders was 23, and for ninth-graders it was 16. What are the chances that:
   a. There will be a true difference greater than 0?
   b. There will be a difference of 5 or more?
   c. The difference will be 3 or less?

   (Solve this problem by using ϵ and also $P.E.$ and compare your results.)

CHAPTER X

CORRELATION

Common use of the term. The correlation concept is used almost daily, if not hourly, in the thinking of both the layman and the statistician. For example, if we say that "fast drivers are likely to have accidents" we mean that there is a correlation between rate of driving and frequency of accidents. Likewise, if we say that bright children do better school work than dull children we mean that there is a correlation between brightness and scholarship. If, in a group of 100 school children, the brightest child had the highest scholastic record, the next brightest had the next highest record, and so on down to the dullest, who had the poorest record of the entire group, we should say that there was very high, in fact perfect, correlation between brightness and scholarship. On the other hand, if we noted that there were many exceptions to the rule that high scholarship accompanied high intelligence, but that in general they tended to agree, we should say that there was a moderate or fair degree of correlation between the two traits.

Definition of correlation. From the examples that have been given the reader can readily see that correlation means the tendency for one factor to vary in some fixed manner in relation to variations in a second factor. Thus, since in general the weight of school children increases as height increases, we say that the two vary together, or are cor-

related. Likewise, since the mental ability of children increases year by year we say that age and mental ability vary together, or are correlated with each other. Furthermore, since the older children tend to be taller and heavier than the younger ones, we say that age is correlated with height and weight. The amount of correlation that exists depends upon the closeness of agreement between the two traits, and is high when there are few exceptions and low when there are many exceptions to the rule that the two variables, traits, or factors, vary together.

Correlation does not necessarily signify causal relationship. As a general rule, when two traits, factors, or variables, such as intelligence and scholarship, are correlated, it is possible to show that one is caused by the other. Thus it would seem perfectly reasonable to assume that intelligence was a causal factor determining school grades. Likewise, a correlation between regularity of attendance and the school marks earned would ordinarily be interpreted as due to a causal influence of attendance upon marks. Such an interpretation is debatable, however, since the good attendance and the high marks may both be due to a third factor, earnestness and seriousness of purpose. Similarly, the correlation between height and weight does not signify that an increase in height causes an increase in weight, nor that variations in weight influence variations in height. All that correlation tells us is that the two factors tend to vary together; the question of causal relation must be determined by the critical analysis of the factors and elements being studied in the particular situation.

Correlation may be of any amount between perfect positive and perfect negative. If the score in one trait, or variable, tends to increase in proportion as the score in the second one increases, we say the correlation is positive. If there are no exceptions whatever to this rule, we say that there is perfect positive correlation. If there is no relationship at all between the two variables, we say there is zero correlation. If, however, the score in one variable goes up as the score in the other goes down, we say that the correlation is negative. If there are no exceptions to this rule in the latter case, and the variations in one are proportional to those in the other, we say there is perfect negative correlation.

To designate these various degrees of correlation numerically it has become the common practice to let perfect positive correlation be represented by +1.00, zero correlation by 0.00, and perfect negative correlation by −1.00. It is seldom, however, when dealing with educational data that one finds a case of correlation exactly at one of these points, hence it is customary to represent intermediate degrees of correlation by such quantities as .95, .72, .51, .17, −.19, −.64, or −.82. All such numbers are called *coefficients of correlation*, and are designated by the small letter ***r***.

Let us illustrate a few cases of positive and negative correlation. If we correlate the time devoted to the study of a lesson with the grade earned on the day's test, we probably will get a positive coefficient of correlation, because the students who study more will tend to get better grades. (Some of them may have read the lesson several times.) If,

however, we have all students read a given passage once and then take a test on finishing it, we probably will get a negative correlation between time required to read it and the test-score earned. This is due to the fact that the slow readers also tend to be poor in comprehension. If we correlated the time required to read the passage with the rate (in words per minute) of reading the passage, we naturally would get perfect negative correlation, or −1.00, because the two items correlated are simply two different (and opposite) ways of expressing the same thing.

Problems: For further practice tell whether you would expect positive or negative correlation in each of the following cases, and why:

1. Intelligence and number of errors on a spelling test.
2. Intelligence and number right on an arithmetic test.
3. Intelligence and time required to solve 15 arithmetic problems.
4. Time required to solve 15 arithmetic problems and the number incorrectly solved.

The scatter diagram as a device to reveal the amount of correlation. Up to this point in this chapter we have been concerned with clarifying the concept of correlation. We shall now consider methods of analyzing the amount of correlation that exists between two variables. One of the most fundamental devices for doing this is the scatter diagram, which in reality is a kind of two-way frequency distribution. It is made by arranging the class intervals for one variable along the left scale and those for the other variable along the top scale, and then placing, in the appropriate square, or *cell*, a mark which will represent each of the two variables for a given individual case.

Let us note in detail how the tabulation of a scatter diagram is made. For example, if the first student in the class represented in Table XXII had a total of 22 ideas written in his paper, and his paper had been estimated as

TABLE XXII. A SCATTER DIAGRAM SHOWING CORRELATION OF .79

(The data correlated are scores on examination papers when scored by counting ideas expressed and when scored by estimating on a 9-point scale.)

Number of ideas	9-point scale estimates								
	1	2	3	4	5	6	7	8	9
21–23					/	///	/	/	/
18–20				/		//	//	/	
15–17				/	/	////	/		
12–14				/	///	/			
9–11			/	////	//	/			
6– 8			/	//		/			
3– 5		/	/						
0– 2	/	/							

(r = .79)

5 on the 9-point scale, we would first find the 21–23 row, just as in making a frequency table, and we would then follow this row to the right until we came to the 5 column, and make our check. If the next case in our table were a student whose score was 11 ideas and a 3-point estimate, we would find the 9–11 row and then go to the right and place a check in the 3 column. This process would be continued until all pairs of scores were checked in their appropriate cells. For accuracy it is ordinarily wise to retabulate, placing a dot over each tally the second time, in order

to be sure that no case has been put in the wrong row or column.

Observing Table XXII, we are impressed by the fact that there is a distinct alignment of the cases diagonally upward and to the right, showing that as scores on one basis of marking increase there is a corresponding increase in the scores on the other basis of marking. In this case we have a positive correlation between number of ideas and scores. If the alignment had been downward and to the right, we should have had a negative correlation. If there had been no alignment at all, and the scores had been scattered promiscuously over the whole surface, we should have had a zero correlation. With a little practice it becomes possible to make a fairly accurate estimate of the amount of correlation between two variables simply by looking at the arrangement of the cases in the scatter diagram.

The mathematical computation of the coefficient of correlation. Although the scatter diagram reveals graphically the amount of correlation that exists between two variables, it does not do so with enough accuracy and precision to be very satisfying to a careful statistician. Furthermore, one cannot always have a scatter diagram handy to show to other people when telling them about the correlation between two traits that have been investigated; consequently there is need for some mathematical process for summarizing into one numerical term the amount of correlation found between the two sets of data. This numerical term is called the coefficient of correlation (r), or the correlation coefficient, and is computed from the scatter diagram such as was illustrated in Table XXII.

TABLE XXIII. ILLUSTRATION OF THE PEARSON PRODUCT-MOMENT METHOD OF COMPUTING THE COEFFICIENT OF CORRELATION FROM A CORRELATION TABLE

(The data are the same as in Table XXII)

No. of Ideas	9-point scale estimates									Computations			
	1	2	3	4	5	6	7	8	9	f	x	fx	fx^2
21–23 *					1 (3)	3 (9)	1 (3)	1 (3)	1 (3)	7	3	21	63
18–20				1 (2)		2 (4)	2 (4)	1 (2)		6	2	12	24
15–17				1 (1)	1 (1)	4 (4)	1 (1)			7	1	7	7
12–14				1 (0)	3 (0)	1 (0)				5	0	0	0
9–11			1 (− 1)	4 (− 4)	2 (− 2)	1 (− 1)				8	− 1	− 8	8
6– 8			1 (− 2)	2 (− 4)		1 (− 2)				4	− 2	− 8	16
3– 5		1 (− 3)	1 (− 3)							2	− 3	− 6	18
0– 2	1 (− 4)	1 (− 4)								2	− 4	− 8	32
f	1	2	3	9	7	12	4	2	1	41		+ 10	168
y	− 4	− 3	− 2	− 1	0	1	2	3	4			(Σfx)	(Σfx^2)
fy	− 4	− 6	− 6	− 9	0	12	8	6	4	5	(Σfy)		
fy^2	16	18	12	9	0	12	16	18	16	117	(Σfy^2)		
fx	− 4	− 7	− 6	− 5	2	14	8	5	3	10	(Σfx)		(Check)
fxy	16	21	12	5	0	14	16	15	12	111	(Σfxy)		

Formula:

$$r = \frac{\Sigma fxy - \dfrac{(\Sigma fx)(\Sigma fy)}{N}}{\sqrt{\Sigma fx^2 - \dfrac{(\Sigma fx)^2}{N}}\sqrt{\Sigma fy^2 - \dfrac{(\Sigma fy)^2}{N}}}$$

Substituting:

$$r = \frac{111 - \dfrac{10 \times 5}{41}}{\sqrt{168 - \dfrac{(10)^2}{41}}\sqrt{117 - \dfrac{(5)^2}{41}}}$$

$$= .79$$

* The number of class intervals here is subject to the same rule as elsewhere, and should ordinarily be 10 or more if the data permit. A smaller number is used here to avoid crowding the page and obscuring essentials.

Steps in the process. The process of carrying out this computation is a bit complex, and is illustrated in detail in Table XXIII. The various steps are described in detail in the following paragraphs.

Step 1, *substituting numbers for tallies.* Comparison of Table XXII with Table XXIII shows that, whereas tally marks were used in the former, figures have been substituted in the latter. This takes away some of the graphic quality of the scatter diagram, but it facilitates the rest of the calculation, particularly if there are very many cases in a cell.

Step 2, totaling the frequencies by rows and columns. The reader will note that there is an f column at the right of the table, just next to the area devoted to the data, and that there is an f row across the bottom. The f column is filled out by adding all cases horizontally across a given row, and the f row is likewise made up of the totals for the several columns.

Step 3, making computations at the right the same as in computing the standard deviation. The reader will note that the columns f, x, fx, and fx^2, with their totals at the bottom, have a very familiar appearance. They are identical with the work which would be done in computing σ (see Chapter VII), the only difference being that the letter x is used instead of the letter d. This is done simply for convenience, in order to distinguish it from y which is used underneath the table. We might use d and d' just as well, were it not for the awkwardness of writing and reciting the formula, and the danger of error due to confusing the two.

Step 4, making computations at the bottom the same as in computing the standard deviation. The rows designated as f, y, fy, and fy^2, under the table, are comparable to the columns designated as f, x, fx, and fx^2 at the right, and also comparable to the familiar f, d, fd, and fd^2 in computing standard deviation. The letter y is used instead of d to distinguish it from x at the right of the table, and the work is arranged horizontally instead of vertically, but otherwise there is nothing new here.

Step 5, multiplying each cell-frequency by its corresponding x value. The reader will notice a number in parentheses in the lower right corner of each cell. These numbers are found by multiplying each cell frequency by the number in the x column in the same row. Thus the number 3 is in parentheses in the top cell of the 5 column. This number 3 is the product of the cell frequency, 1, times its x deviation, 3. Similarly, the cell to its right has the number 9, the product of the cell frequency 3 by the x deviation 3. This process is carried out for the remaining cells in the same way.

Step 6, adding up the numbers in parentheses by columns. Since the numbers in parentheses are products of frequencies times deviations, they may be represented by the symbol fx, just as truly as the products in the fx column at the right of the table. Hence we have an fx row near the bottom of the table, arrived at by adding up all the numbers in parentheses by columns. Thus in the 1 column we have only one number in parentheses, −4. Hence we have the entry −4 in the fx row underneath. The next column has

entries of -3 and -4 in parentheses, and we have the total -7 in the fx row underneath. This process is continued for the other columns in the same fashion, and the algebraic total of the fx row, which is 10, is written at the end of the row, along with the totals for the fy and fy^2 rows.

Step 7, checking the work up to this point. Since the total of fx should be the same whether it is found by rows or by columns, any failure of the two items to agree is an indication that an error has been made somewhere in the work. In the present case Σfx is 10 in each case, and we are assured that the work is accurate so far. There is one limitation of this check. It does not apply to the items Σfx^2, Σfy or Σfy^2. An error may be made in these items even though the Σfx items are correct.

Step 8, multiplying fx by y. After checking the accuracy of the fx computations the next step is to get the product fyx for each column. This is done by multiplying each fx in the row toward the bottom of the table by the y value three lines above it. Thus the first fx entry at the left is -4, and the y entry above it is also -4, giving the product 16. The next fx entry is -7 and its y value is -3, giving the product 21, and so on across. The total of all these products of fx and y is designated Σfxy, and in the present case is 111.

Step 9, substituting in the formula. The formula for the correlation coefficient is written in numbers of different ways by different statisticians, but all yield the same numerical result in the end. The formula given below Table XXIII seems preferable because of its convenience,

even though it appears a bit complex. Its derivation or origin cannot be considered in an elementary book on statistics. The reader will simply have to take for granted that this formula is correct, or else consult a more advanced treatise for a mathematical explanation of it. To substitute the desired items in the formula should be very simple, since each is clearly labeled in the table.

Step 10, *clearing up the complex fraction.* Once the numbers are inserted in the formula there remains only the routine arithmetical work of clearing up the fraction. This work is considerably speeded and lightened by the use of the slide rule, table of square roots, or logarithms. Certainly there should be no excuse for doing the square roots by the usual long method.

How to interpret the correlation coefficient in terms of amount of relationship between the variables. The following question is certain to arise: "Just how much correlation is there when the coefficient is .79?" Is there 79 per cent of a perfect correlation? Do 79 people out of 100 reveal correlation and the other 21 not? No such simple answers can be truthfully given. Various efforts at a "layman's interpretation" of the correlation coefficient have been made, usually with little success. The safest and most dependable interpretation of the coefficient .79 in the present case is that it represents the amount of agreement shown in the scatter diagram in Table XXII. Increased experience and increased familiarity with correlation through working with data manifesting it in different degrees will gradually give a fairly distinct "feeling" for a coefficient

of a given size as representing a scatter diagram of a definite appearance.

Some readers will not be content with the above basis for interpreting correlation coefficients, particularly in the early stages of their statistical experience, and will wish some more tangible basis for judging their size and importance. Thus they may get some comfort or help from such an empirical classification as that given by Rugg,[1] which we may summarize as follows:

> Below .15 or .20, "negligible or indifferent."
> From .15 or .20 to .35 or .40, "present but low."
> From .35 or .40 to .50 or .60, "markedly present."
> Above .60 or .70, "high."

The importance that is attached to a correlation coefficient also depends upon the nature of the items correlated, and the educational situations in which they are to be applied. For example, a correlation of .50 may be sufficient basis for a general policy, but not high enough to apply in dealing with individual school pupils. Thus a coefficient of .50 might show that intelligence was a factor in college success, but would not be high enough to warrant excluding from college particular students who made low scores on the intelligence test. A coefficient of .65 between intelligence and success in a specific subject might be very significant, whereas the same coefficient as an index of the reliability of a standardized test would be interpreted as unsatisfactory. It is well to remember that two factors greatly

[1] H. O. Rugg, *Statistical Methods Applied to Education*, p. 256. Boston, Houghton Mifflin Company, 1917.

influence the size of coefficients, the accuracy of the original measurements and the degree of selection of the cases used in the computation; the greater the accuracy of the original measurements, the higher the coefficient, and the narrower the range represented (i.e., the higher the selection) in either or both of the variables whose relationship is being determined, the lower will be the coefficient obtained.

Some methods of computing correlation without the scatter diagram and correlation table. The method of computing correlation coefficients which we have just learned is the best and the shortest when the number of pairs of scores is large and when the scores are large numerically. Numerous substitutes are available, however, which will yield the same or practically the same results, and which may be easier to learn and more economical to calculate, especially when the cases are not numerous and the scores are not large numerically. We should beware of trying to learn several methods without learning any of them well, and for that reason many users of this book had best omit the methods next described. For the service of special groups of readers, however, we give two of the more common special methods, namely, the Pearson product-moment method for ungrouped pairs and the Spearman rank-differences method.

The Pearson method for ungrouped pairs. A comparatively simple method of computing correlation is shown in Table XXIV. Nothing is required except squaring the scores and multiplying the scores within each pair. The totals of the columns thus filled out are ready to insert in

TABLE XXIV. ILLUSTRATION OF THE COMPUTATION OF THE PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT FROM UNGROUPED PAIRS OF DATA

Pupils	Scores in Test 1 (X)	Scores in Test 2 (Y)	Products of scores (XY)	Squares of scores on Test 1 (X^2)	Squares of scores on Test 2 (Y^2)
(1)	(2)	(3)	(4)	(5)	(6)
1	3	6	18	9	36
2	5	8	40	25	64
3	4	7	28	16	49
4	9	11	99	81	121
5	11	12	132	121	144
6	4	7	28	16	49
7	2	4	8	4	16
8	5	9	45	25	81
9	7	10	70	49	100
10	8	12	96	64	144
11	10	8	80	100	64
12	2	6	12	4	36
13	9	12	108	81	144
14	6	8	48	36	64
15	8	9	72	64	81
15 (N)	93 (ΣX)	129 (ΣY)	880 (ΣXY)	695 (ΣX^2)	1193 (ΣY^2)

Formula: *

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

Substituting:

$$r = \frac{880 - \frac{93 \times 129}{15}}{\sqrt{695 - \frac{(93)^2}{15}}\sqrt{1193 - \frac{(129)^2}{15}}}$$

$$= .80$$

* This formula is taken from Kelley, T. L., *Statistical Method*, p. 164.

the formula. The squaring, multiplying, and adding may become very laborious if the scores are large numerically, or if there are very many pairs of scores. Thus if we were correlating intelligence quotients for 150 children with their scores on a spelling test which included over 100 words, we should have to multiply and square many numbers involving three figures, and then add their huge squares and products.

A device for eliminating a part of this arithmetical labor is to subtract the smallest X score from all the other X scores, and the smallest Y score from all the Y scores, and then correlate these much smaller remainders. These reduced scores are still in the same rank or order as they were before, and hence the correlation coefficient will be exactly the same. Applying this device in Table XXIV, we could reduce all scores on Test 1 by 2 points and all on Test 2 by 4 points without being troubled by negative numbers. But since our numbers are already very small, it would not be worth while to go to the trouble in this case.

The Spearman rank differences method. The method illustrated in Table XXV involves ranking the scores and then squaring the differences between the ranks for each pair of scores. It sometimes happens that the original data are in the form of ranks instead of actual scores, in which case this method is especially desirable.

Table XXV presents, in columns 2 and 3, the scores made by fifteen students in two applications of a spelling test. In column 4 is presented the rank achieved by each student in Test 1. Student number 5 spelled 40 words correctly, the highest number of correct spellings made, thus entitling him to rank 15. Notice that students numbered 4, 8, and

TABLE XXV. ILLUSTRATION OF THE COMPUTATION OF THE COEFFICIENT OF CORRELATION BY THE SPEARMAN RANK DIFFERENCES METHOD [1]

(The data are from two spelling tests.)

Students	Scores in Test 1	Scores in Test 2	Ranks in Test 1	Ranks in Test 2	Differences in ranks (D)	Differences squared (D^2)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	26	38	6	13	7	49
2	29	32	9	8	1	1
3	34	40	11	15	4	16
4	38	35	13	10	3	9
5	40	39	15	14	1	1
6	23	19	4	3	1	1
7	27	30	7.5	7	0.5	0.25
8	38	34	13	9	4	16
9	27	24	7.5	5	2.5	6.25
10	33	37	10	12	2	4
11	25	26	5	6	1	1
12	21	20	3	4	1	1
13	15	18	1	2	1	1
14	38	36	13	11	2	4
15	19	17	2	1	1	1
$N = 15$						$\Sigma D^2 = 111.5$

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$
$$= 1 - \frac{6 \times 111.5}{15\,[(15)^2 - 1]}$$
$$= .80$$

14, all obtained 38 correct spellings, and that rank 13 has been assigned to each. In ranking scores for rank differences correlation computations, whenever two or more students have identical scores, each is assigned the midpoint of the range of ranks which they would have occupied if scores had not been identical. Thus, if each of these students (numbers 4, 8, and 14) had had different scores,

[1] This method is discussed and evaluated in Rugg, H. O., *Statistical Methods Applied to Education*, pp. 284–86.

and yet all of them were higher than 34 (which receives rank 11) they would have occupied ranks 12, 13, and 14. Thirteen is the midpoint of this range of ranks, hence it is assigned to all. Note also that students numbered 7 and 9 also have identical scores. Student number 2 received rank 9, and students numbered 7 and 9 have scores next below him. If they had not been identical and yet were the next in terms of size, they would have used up ranks 7 and 8. However, the scores are identical, so the midpoint of this range of ranks, or 7.5, is assigned to each.

In column 5 are presented the ranks made by these students in test 2. It so happens that there were no tied scores in the second test, so the matter of ranking was comparatively simple.

In column 6 of Table XXV are presented the differences in the ranks obtained by students on the two tests. After some practice, this column may generally be omitted, but for the beginner it is better to retain it. In column 7 are the squares of the rank differences found in column 6. The sum of the differences squared (ΣD^2), presented in column 7, is 111.5. The number of pairs of measurements is 15. With this information obtained, these sums may be substituted in the formula and computation of the coefficient of correlation completed.

The symbol representing the correlation coefficient computed by the rank differences method is the Greek letter rho, ρ. The formula for computing it is:

$$\rho = 1 - \frac{6\,\Sigma D^2}{N(N^2 - 1)}$$

The remaining steps in the process are: (1) finding the squares of the differences in rank (D^2); (2) obtaining the sum of these squared differences (ΣD^2); (3) multiplying this sum by six; (4) dividing the result obtained in step 3 by the quantity $N(N^2 - 1)$; (5) subtracting the quotient obtained in step 4 from 1.

In concluding this brief treatment of the rank-differences method of correlation, special emphasis should be laid upon the fact that this method of correlation deals with *rank* differences, and not *score* differences. Most of the errors made by the beginner probably will result from attempting to work with score differences, rather than rank differences.

The coefficient ρ obtained by the rank-differences method is not identical with r obtained by the Pearson product-moment method, but ordinarily varies from it by only a few points. Since the method is used only for small groups there is little justification for extreme accuracy, and the few points difference are of little importance.

Summarization of steps in the rank-differences method. The computing of a coefficient of correlation by the rank-differences method may be summarized in the following steps:

1. Set down the pairs of scores in adjacent columns (as in columns 2 and 3, Table XXV).

2. Rank each set of scores (as in columns 4 and 5, Table XXV).

3. Obtain the differences in rank of each of the pairs of measurements (as in column 6, Table XXV).

4. Set down the squares of the rank differences (as in column 7, Table XXV).

5. Get the total of the squares of the rank differences, (ΣD^2) at the bottom of column 7 in Table XXV).

6. Substitute the sum of the squares and the number of pairs of measurements in the formula and solve for ρ.

Reliability of correlation coefficients. In an earlier chapter we discussed reliability of various statistical quantities, and pointed out the great importance of computing the reliability of any statistical quantity before putting faith in its significance. The treatment of reliability of correlation coefficients was omitted in that chapter because correlation had not yet been presented. However, the reliability of each coefficient of correlation should be computed as an integral part of the computation. This is of such importance that it is given a considerable portion of the following chapter, where it will be dealt with at length and in considerable detail.

TRUE-FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. Most of the calculations for obtaining the correlation coefficient are the same as those for computing standard deviation.
2. To correlate the test scores of 20 boys with those of 20 girls would be absurd.
3. If the two fx items check, that is a guarantee of the accuracy of all the work.
4. A coefficient of .58 indicates the same degree of relationship in one type of data as in another.

5. Correlations based on over 100 cases are of doubtful worth, because individual cases are lost in the mass.
6. It is more difficult for two statisticians to get the same correlation from given data than for them to get the same mean.
7. There is less need of computing reliability for correlation coefficients than for other statistical quantities.
8. When a high correlation is established between two variables we know that one causes the other.
9. Making a scatter diagram is the equivalent of tabulating two frequency tables simultaneously.
10. The correlation coefficient shows the per cent of relationship involved.
11. It makes no difference which variable is written across the top and which down the left side in a correlation table.
12. A positive correlation shows that as one factor increases the other tends to increase.
13. No use should be made of a correlation coefficient if it is below .50.
14. Correlation may be defined as the nature and extent of quantitative relationship between two or more variables.
15. Correlation of the scores for two groups is a good way to find the reliability of the difference between the groups.
16. It is impossible to compute correlation without making a scatter diagram.
17. It is essential that the number and size of class intervals for both variables be the same.
18. A negative correlation shows that as one factor decreases the other tends to decrease.
19. If a correlation coefficient is negative it is useless for making predictions.
20. The rank-difference method of getting correlation is best whenever the cases number more than twenty.
21. The Pearson method of ungrouped pairs enables the statistician to avoid fractions in computing the coefficient of correlation.

TABLE XXVI. DATA TO BE USED FOR PRACTICE IN COMPUTING COEFFICIENTS OF CORRELATION

(The data are from the official records of a college registrar, and are for a part of the freshman class.)

Pupil	High-school units	English test errors	English grade	Intelligence	Pupil	High-school units	English test errors	English grade	Intelligence
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
1	16	37	4	35	31	$18\frac{1}{2}$	26	4	75
2	16	63	–	31	32	15	34	5	28
3	16	26	6	42	33	17	29	3	28
4	18	18	5	88	34	17	21	5	75
5	15	9	6	31	35	$16\frac{1}{2}$	37	5	31
6	16	25	5	16	36	18	20	5	91
7	$16\frac{1}{2}$	63	5	15	37	$15\frac{1}{2}$	48	1	10
8	$17\frac{1}{2}$	19	4	60	38	16	33	4	49
9	$15\frac{1}{2}$	56	3	11	39	17	28	5	42
10	$16\frac{1}{2}$	5	6	90	40	13	50	2	16
11	18	19	4	16	41	15	23	6	88
12	17	44	3	11	42	15	18	5	81
13	15	29	1	22	43	18	23	5	58
14	$17\frac{1}{2}$	23	5	47	44	17	24	2	12
15	15	39	4	51	45	$16\frac{1}{2}$	20	4	60
16	16	30	4	26	46	$18\frac{1}{2}$	24	1	13
17	$18\frac{1}{2}$	28	4	68	47	$16\frac{1}{2}$	22	3	26
18	16	11	5	76	48	17	45	1	23
19	$15\frac{1}{2}$	10	6	90	49	16	32	5	75
20	$17\frac{1}{2}$	31	4	72	50	16	46	2	49
21	$17\frac{1}{2}$	12	5	58	51	$16\frac{1}{2}$	42	4	47
22	17	18	5	57	52	18	22	4	68
23	17	40	–	48	53	17	41	4	57
24	$16\frac{1}{2}$	62	4	9	54	17	28	6	42
25	17	16	5	48	55	16	20	5	95
26	$15\frac{1}{2}$	36	3	68	56	18	25	3	31
27	16	40	4	45	57	19	69	1	23
28	$17\frac{1}{2}$	34	4	12	58	$15\frac{1}{2}$	24	3	–
29	16	–	4	30	59	17	39	4	45
30	$17\frac{1}{2}$	24	4	38	60	16	62	1	16

QUESTIONS AND EXERCISES

In Table XXVI are actual data from a registrar's records regarding a few college freshmen, chosen arbitrarily because of their position in the alphabetical list. It will be noted that information is lacking for certain pupils on some of the items, in which case there is a dash. In computing the correlation coefficient for any given pair of items the existence of a blank in one item will require the elimination of the data for that pupil on the other item with which it is being correlated. Make the following computations and answer the following questions:

1. What is the correlation between number of high-school units offered for admission and grades earned in English?
2. Do the brighter students tend to offer more high-school units than the duller ones? What correlation?
3. Is the intelligence test a better prediction of English grades than the special English test? What is the difference in correlation?
4. How does the fact that the English test shows a negative correlation with English grades affect its usefulness in predicting a student's success in freshman English?
5. Do bright students make fewer errors on the English test than dull ones? How much correlation is there?
6. How much correlation is there between the number of high-school units offered and the number of errors made on the English entrance test? What would this correlation signify?
7. Compute the correlation for each of the problems mentioned above, using only the first fifteen students and employing the rank-differences method. Compare with the coefficients for the entire number by the previous method.
8. Compute the above-mentioned correlations for the first 15 students, using the Pearson method for ungrouped pairs and compare with those secured by the rank-differences method.

CHAPTER XI

FURTHER ASPECTS OF CORRELATION

In the last chapter we explained the meaning of correlation, and showed how to compute the ordinary correlation coefficient. There are several other aspects of correlation, however, which the beginning student of statistics and the ordinary teacher or school administrator should at least know about in order to be an intelligent reader of statistical research reports. The first of these aspects of correlation should be learned to the point of actually acquiring skill in computation and use in research. The others may merely warrant an understanding and acquaintance. Each of these will be explained in the present chapter.

Reliability of the coefficient of correlation. In a previous chapter the matter of reliability of statistical quantities has been explained and illustrated. The reliability of the correlation coefficient, r, was not dealt with in that chapter because the whole question of correlation was still unfamiliar at that point. We are now ready to carry on the treatment of reliability that was begun there, and to apply it to the correlation coefficient.

If we study a group of 50 school children as regards attendance and marks, and find a correlation of .29 between these two variables for this observed sampling, how safe are we in generalizing from this observation and saying that children in general will reveal a positive correlation between attendance and marks? In other words, how reliably can

we predict, on the basis of the correlation in 50 observed cases, the correlation we should expect to find if we were to study all the school children in the United States, of which these 50 are considered to be a representative sampling? Until we have a device for answering such questions as this we are not in a position to apply our conclusions outside of the restricted group upon which we have made actual observations.

Reliability depends on two factors. Clearly, the reliability of such a generalization from observed cases to the universe of which they are a representative sampling depends upon two factors:

1. The size of the correlation coefficient. If the correlation is extremely high, it is improbable that another sampling of 50 cases would manifest a zero or negative correlation.

2. The number of cases included in the observed sampling. Chance fluctuations are not so likely to be important influences in determining the correlation in a group of 5000 students as in a group of 5 students. A correlation coefficient based on only 5 students would be worthless because one or two unusual cases might have a very great influence on the coefficient, while one based on 5000 cases would probably vary only very slightly from that derived from 5000 additional cases.

Since, therefore, the reliability of the correlation coefficient depends on the size of r and the size of N, the formulas for determining its reliability make use of both of these

quantities. Without going into mathematical derivations of the formulas here, we shall simply state that they are as follows:

$$\epsilon_r = (1 - r^2)/\sqrt{N}$$
$$\text{Ratio} = r/\epsilon_r$$

(For chances that the true r is above zero, see Table XIX).

The meaning of these formulas. Let us examine these formulas in detail and interpret their meaning. The first one gives the standard error of the correlation coefficient. Thus, if we take the case of a correlation of .29 between attendance and marks, with 50 students in the group, we have

$$\epsilon_r = \frac{1 - (.29)^2}{\sqrt{50}} = .13$$

The ϵ_r, .13, tells us that if we should compute r from an infinite number of samplings of students, under conditions similar to those prevailing in the present study, the chances are 2 to 1 (since ϵ takes in the middle two thirds of the cases) that r would not go above $.29 + .13$ nor below $.29 - .13$. Thus the chances are 2 to 1 that the true correlation lies somewhere between .16 and .42.

What, then, are the chances that the true correlation, based on an unlimited number of cases, would be above zero? The second formula enables us to answer this question. If we find how many ϵ_r units r is above zero, we can readily discover what the chances are that chance variations due to sampling could cause the true correlation to fall on or below zero. Thus, in the present case the zero point

is .29 points below the observed r. Since ϵ_r is .13, the zero point is .29/.13, or about 2.2 ϵ_r below the observed r.

By consulting Table XIX in Chapter IX, we find that the chances of a given case's falling within 2.2 ϵ_r, units of the central tendency are 70.9 to 1. In other words, the chances are 71 to 1 that if we observed an unlimited number of cases we would get a correlation somewhere above zero. This statement of chances is the most useful indication of reliability which we can get.

Increasing reliability. How, then, can the reliability of a correlation coefficient be increased? The answer is, by increasing the number of observed cases. Thus if we had a correlation of .29 based on 100 cases instead of 50 cases, we should find ϵ_r to be .09 instead of .13, and thus our r of .29 would be .29/.09, or 3.2 ϵ_r units above zero, with the chances 1450 to 1 that the true correlation is above zero. It is obvious, therefore, that an increase in the number of cases in the group (N) will greatly increase the reliability of the coefficient of correlation. It should be borne in mind, however, that the r which results from including additional cases will not necessarily be the same as the one first obtained. It is, however, as likely to be higher as to be lower.

Reliability of the difference between two correlation coefficients. Quite often there is a need for comparing two tests, teaching procedures, or other items on the basis of the extent to which they correlate with some known criterion. Thus a university had collected data regarding the freshman class and found that the English test given on

entering the university correlated .37 with the scholarship averages in all subjects during the first semester. The correlation between the intelligence test and the scholarship averages was .44. There was a difference of .07 between the two coefficients, favoring the intelligence test as being the better instrument for predicting freshman scholarship. How important was this difference? What were the chances that other freshman classes in future years would manifest a higher correlation in the case of the intelligence test than in the English test? The reliability of this difference is to be determined by the same type of procedure as was learned in Chapter IX for finding the reliability of a difference between two means. It depends upon the reliability of the two correlation coefficients that are being compared.

Clearly a difference between two highly reliable coefficients will be much more reliable than a difference between two coefficients each of whose reliability is low. And, clearly also, the difference is much more reliable if it is large than if it is small. In the present case the reliability of each of the coefficients compared is fairly high, because the number of cases is large, there being 486 students included in the study. (Find ϵ_r for each of the coefficients, .37 and .44).

Formulas for use. The formulas for determining the reliability of the difference are as follows:

$$\epsilon_D = \sqrt{(\epsilon_{r_1})^2 + (\epsilon_{r_2})^2}$$

$$\text{Ratio} = D/\epsilon_D$$

(For chances that the true D is above zero, see Table XIX, page 137.)

In the present case, ϵ_{r_1} is .039 and ϵ_{r_2} is .037. We substitute in the formula and get:

$$\epsilon_D = \sqrt{(.039)^2 + (.037)^2}$$
$$= .054$$
$$\text{Ratio} = .07/.054$$
$$= 1.3$$

(Chances, by Table XIX, page 137, 9.3 to 1 that there is a true difference in favor of the intelligence test.)

Since the difference, .07, only proves to be 1.3 times its ϵ_D of .054, we see by Table XIX that the chances are only 9.3 to 1 that the intelligence test is really a better predictive instrument than the English test. This is not a high enough ratio for us to be certain of the superiority, since, as we learned in Chapter VIII, the normal curve does not by any means approach the base line at a distance of 1.3 ϵ units from the mean. The ratio should be well up toward 3 (McCall uses 2.78 as the point) in order to yield what is called "practical certainty" that the difference is real rather than due to chance factors of sampling.

Probable error as a unit for measuring the reliability of correlation coefficients. The same considerations which were discussed in Chapter IX regarding the relative merits of ϵ and *P.E.* apply in the present connection. We need do no more here, therefore, than to give the necessary formulas, because the logic of the work is the same as for what we have just discussed.

The formulas for the reliability of a correlation coefficient are:

$$P.E._r = .67449\,\epsilon_r$$
$$\text{Ratio} = r/P.E._r$$

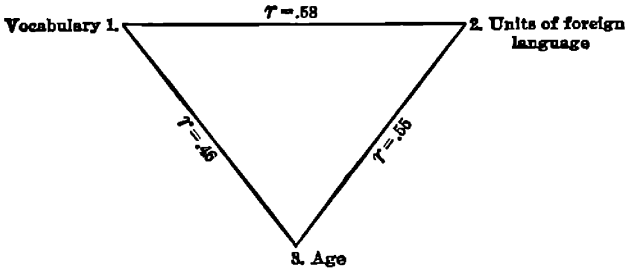
(For chances that the true correlation is above zero, see Table XXI, pages 148–49.)

The formulas for determining the reliability of a difference between two correlation coefficients are:

$$P.E._D = \sqrt{(P.E._{r_1})^2 + (P.E._{r_2})^2}$$
$$\text{Ratio} = D/P.E._D$$

(For chances that the true difference is greater than zero, see Table XXI.)

Partial correlation. Sometimes a correlation that is found to exist between two variables is suspected of being due to the influence of some third factor, rather than to any real relationship between the two themselves. There is need, therefore, for some method of holding constant such a third factor in order to get the true relationship between the two being studied. Thus a correlation of .58 was found to exist between size of vocabulary and number of units of foreign languages studied. Some one observed, however, that the older students in the group seemed to have higher scores for vocabulary as well as more units of foreign languages, and suggested that possibly the correlation of .58 might be due to the influence of the age factor. All three factors were intercorrelated, therefore, with the results shown in Figure 16. (It will be noted that the correlation coefficient for a given pair of variables is written on the line connecting them.) Substituting in the formula, the final result was a coefficient of .44 instead of the original .58. Thus it would appear that the age factor really had been responsible for a part of the original correlation between the two variables, since its elimination caused a drop of 14 points in the correlation coefficient. To be certain of this we should compute the reliability of the difference of .14 between the two coefficients, as previously explained.



Formula:

$$r_{12\cdot 3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1 - (r_{13})^2} \quad \sqrt{1 - (r_{23})^2}}$$

Substituting:

$$r_{12\cdot 3} = \frac{.58 - (.46)(.55)}{\sqrt{1 - (.46)^2} \quad \sqrt{1 - (.55)^2}} = .44$$

FIG. 16. ILLUSTRATION OF THE STEPS IN FINDING A PARTIAL CORRELATION COEFFICIENT

The problem is to find the correlation between amount of foreign language study and size of vocabulary, with age held constant.

It is possible for a coefficient to be either increased or decreased in size where a third factor is held constant, depending upon the size of the three intercorrelations and whether some of them are positive or negative.

It is possible, also, to hold several factors constant, as compared with the one factor held constant in the above example. The process becomes rather complex, however, and is beyond the scope of this book.[1]

Multiple correlation.[2] Just as two or more vague and uncertain clues to a crime may, when put together by the

[1] McCall, *How to Experiment in Education*, pp. 233–40, gives formulas up to seven variables.

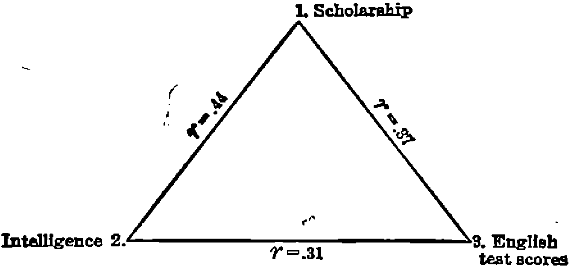
[2] For a good treatment of multiple correlation, see Buckingham, B. R., "Multiple Correlation," *Journal of Educational Research*, IX, February, 1924, pp. 149–54.

detectives, shed distinct light on the puzzle when neither alone was very convincing evidence, so two or more small correlation coefficients may sometimes be combined in such a way as to be of the utmost service when neither one alone was of very great importance. Thus if we wished to predict a college freshman's most probable scholastic achievement from a knowledge of his intelligence, which in previous years had been found to correlate only .44 with scholarship, we might hesitate to put great faith in the prediction. Still less would we have faith in a prediction from his English entrance test, which showed only a correlation of .37 with scholarship. By multiple correlation, however, it is possible for us very quickly to determine the possibility of improving the accuracy of our prediction in case we should combine the intelligence and English test scores in the optimum proportions by suitable weighting.

In other words, multiple correlation is useful when each of two variables has been correlated with a third, and we wish to know whether we could get a higher correlation if we were to combine the two variables and correlate the composite with the third variable. Going back to our illustration, if intelligence correlates .44 with scholarship and English scores correlate .37, multiple correlation will tell us whether a composite of intelligence and English scores will yield a correlation higher than .44. This will be clear if we examine Figure 17.

Substituting the respective coefficients in the formula, we find that it would be possible to get a coefficient of .51 if we were to weight the intelligence scores and the English test

scores in the optimum proportions and correlate the composite result with scholarship. This does not mean that simply adding each pupil's intelligence and English test



Formula:

$$R_{1\cdot 23} = \sqrt{\frac{(r_{12})^2 + (r_{13})^2 - 2(r_{12})(r_{13})(r_{23})}{1 - (r_{23})^2}}$$

Substituting:

$$R_{1\cdot 23} = \sqrt{\frac{(.44)^2 + (.37)^2 - 2(.44)(.37)(.31)}{1 - (.31)^2}} = .51$$

FIG. 17. ILLUSTRATION OF THE STEPS IN FINDING A MULTIPLE CORRELATION COEFFICIENT

The data are simple intercorrelations between the three variables, scholarship, intelligence, and English test scores.

scores would make it possible to get a coefficient of .51. We should have to apply another formula in order to discover the proper weights to give each of the two scores to be combined. Since the purpose of the present discussion is to show what multiple correlation is, rather than to develop skill in using it, we shall not go into the details of weighting here. A person who has need for this should study the topic further in a more advanced work.

It is possible to tell fairly well by simple inspection whether the multiple correlation coefficient will be higher

than the highest simple correlation coefficient. If the correlation between variables 2 and 3 is distinctly lower than that between each of these and variable 1, there will be an increase. Thus, in Figure 17, the correlation between variables 2 and 3 is .31 when these variables correlate .44 and .37 respectively with variable 1. Hence the authors knew beforehand that $R_{1 \cdot 23}$ would be somewhat larger than .44. If r_{23} had been .55 we should have known that it would be worthless to compute $R_{1 \cdot 23}$ because there would be no hope of getting a significant increase above .44.

When is the increase significant? One question remains before leaving multiple correlation. That is, when do we know that the increase in the correlation coefficient made possible by multiple correlation is large enough to be significant? Thus, in the above example, we had a correlation as high as .44 already, but found that we could get .51 by making the combination. Is this gain of 7 points enough to deserve any special consideration, or could it be due to mere chance?

In answering this question the first thing to consider is that the multiple correlation coefficient cannot possibly be less than the highest simple correlation coefficient upon which it is based. If it is, there is an error in the work. Thus, in the present case $R_{1 \cdot 23}$ could not possibly be less than .44 except through an error in the work. There will ordinarily be a slight increase, due to purely chance factors, and this increase should never be taken seriously until its reliability has been determined, as was explained in an earlier part of this chapter.

In determining the reliability of the present increase of 7 points from .44 to .51, we substitute in the formulas previously given for reliability of a difference between two correlation coefficients:

$$\epsilon_D = \sqrt{(\epsilon_r)^2 + (\epsilon_{R_{1\cdot 23}})^2}$$

This necessitates finding ϵ_r and $\epsilon_{R_{1\cdot 23}}$ by means of the formulas:

$$\epsilon_r = (1 - r^2)/\sqrt{N}$$

$$\text{and} \quad \epsilon_R = (1 - R_{1\cdot 23}{}^2)/\sqrt{N}$$

Thus, since N in each case is 486, and r is .44 and $R_{1\cdot 23}$ is .51, we find ϵ_r to be .03 and $\epsilon_{R_{1\cdot 23}}$ to be .04. Substituting these in the formula, we have:

$$\epsilon_D = \sqrt{(.03)^2 + (.04)^2}$$
$$= .05$$
$$\text{Ratio} = .07/.05 = 1.4$$

Thus we find that the new coefficient .51 is only 1.4 ϵ_D units higher than the original coefficient of .44, with chances only 11.4 to 1 that there is a real difference. Since it should be about 3 ϵ_D units higher in order for us to be certain that it is not due to chance, we conclude that it would not be worth while in this case to combine the two sets of data into the composite.

The use of *P.E.* in connection with partial and multiple correlation coefficients. The method which we have just studied for determining whether a multiple correlation coefficient is significantly higher than the previous simple correlation coefficient can be be carried out in terms of *P.E.* just as well as in terms of standard error. Likewise,

when partial correlation has been used to hold a third variable constant, the difference between the original and the partial correlation coefficient can be tested for its significance by use of *P.E.* as well as ϵ.

The formulas are the same as we have already studied, except in insignificant details. Thus, to get the reliability of a difference between a multiple and a simple correlation coefficient, the formulas are:

$$P.E._D = \sqrt{(P.E._r)^2 + (P.E._{R_{1\cdot 23}})^2}$$
$$\text{Ratio} = D/P.E._D$$

(For chances that the true difference is above zero, see Table XXI.)

The formulas for the reliability of the difference between a partial and a simple correlation coefficients are:

$$P.E._D = \sqrt{(P.E._r)^2 + (P.E._{r_{1\cdot 23}})^2}$$
$$\text{Ratio} = D/P.E._D$$

(For chances that the true difference is more than zero, see Table XXI.)

Curvilinear correlation. All our discussion of correlation up to this point has been based on the assumption of a straight-line relationship. In other words, we have been considering "rectilinear" correlation. We should at least be aware of the fact, however, that there is a different type of correlation in which the relationship is curved, rather than straight. In other words, we should get acquainted with the concept of "curvilinear" correlation.

Let us take an example. In correlating the size of schools and the per capita cost of education we should expect the cost per pupil to increase as the school enrollment decreased. It is plain, however, that when the school gets extremely

small the cost per pupil begins to mount up more rapidly than before. Thus there is a relationship all along, but it is not in a fixed proportion throughout the entire range of sizes of schools. This point will be further clarified by means of Figure 18, which is drawn from imaginary data, and in which there is purposeful exaggeration of the curvilinearity in order to emphasize the principle or concept.

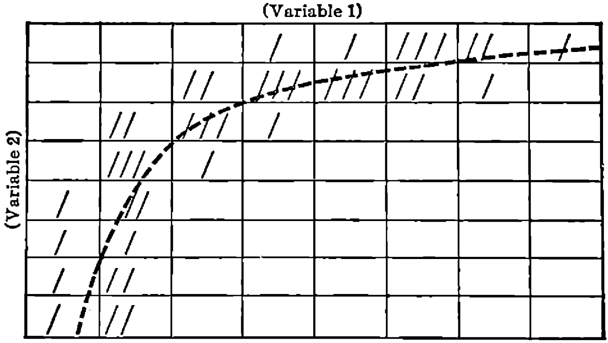


FIG. 18. ILLUSTRATION OF THE MEANING OF CURVILINEAR CORRELATION

The data are purely imaginary, and purposely exaggerate the element of curvilinearity for greater emphasis.

The mathematical computation of curvilinear correlation coefficients is a bit complex, making use of the work that is done in computing the rectilinear coefficient, and adding a number of additional steps. Details of its computation are omitted here because of the elementary scope of this book, and because of the relative infrequency of its use in ordinary educational statistics.

There is, however, an approximation method of computing the amount of curvilinearity, which can be used by any one who knows how to compute ordinary straight-line cor-

relation. It consists of dividing the correlation table up into three or more parts by horizontal lines, computing r for each of these parts, and then averaging the three or more r's thus derived. This amounts, in reality, to getting a "broken-line correlation" rather than a strictly curved-line one, but the broken line is a closer approximation to the true line of relationship than is a single straight line.

The question of when to go to the trouble to compute the curvilinear coefficient is answered in two ways:

1. Compute it if the scatter diagram reveals a noticeable curvilinearity as regards the arrangement of the tallies.

2. Compute it if the analysis of the particular problem, data, or situation suggests that there is good reason to expect a curvilinear relationship.

The curvilinear correlation coefficient is by nature either the same size as the rectilinear coefficient or larger than it, but never smaller. Any chance curvilinearity will raise the coefficient slightly. This fact misleads some beginners, and tempts them to compute the curvilinear coefficient in order to get as high a figure as possible. One should never attach any particular significance to the difference between the rectilinear and the curvilinear coefficients until he has computed the reliability of the difference between them in the manner previously explained and illustrated.

TRUE-FALSE TEST

Decide whether each of the following statements is true or false and be prepared to defend your decision.

1. Partial correlation, as its name implies, is only a halfway measure, and should be used only for approximate results.

2. Computing ϵ_r is a substitute for the labor of actually measuring or observing a large number of additional cases.
3. To find a multiple correlation coefficient requires squaring a number of simple correlation coefficients.
4. The formula for ϵ_r is $(1 - r^2)/\sqrt{N}$.
5. Most educational data manifest straight-line correlation.
6. Partial correlation cannot be used to hold more than one variable constant.
7. If r is higher than .70, it cannot possibly be due to chance.
8. High correlation, like high variability, tends to reduce reliability.
9. When two variables increase or decrease at identical rates a curvilinear correlation should be computed.
10. Multiple correlation tells the relative weights to give to the several tests in order to mark a pupil in a subject.
11. The partial correlation coefficient is always higher than r.
12. Multiple correlation tells us whether a composite score would give a higher prediction than a score from a single test.
13. A correlation coefficient is a device for generalization.
14. The multiple correlation coefficient cannot possibly exceed r, and is likely to fall below it.
15. The standard error of the difference between two correlations is computed in a manner similar to the standard error of the difference between two means.
16. Two correlation coefficients should be compared on the basis of the reliability of their difference.
17. The interpretation of the probable error of r is similar to that of the probable error of other statistical quantities.
18. Partial correlation enables us to hold one factor constant in order to determine the correlation between two others.
19. Additional observed cases will tend to increase the reliability of the correlation coefficient.
20. When a partial correlation coefficient is higher than r, the significance of the difference is determined by computing the standard or probable error of the difference.

QUESTIONS AND EXERCISES

In Table XXVII are data regarding a group of 64 summer-school students in a university, consisting of simple intercorrelation coefficients. Find the answers to the following questions:

TABLE XXVII. DATA FOR PRACTICE IN CONNECTION WITH SPECIAL ASPECTS OF CORRELATION

(The data are simple intercorrelation coefficients for 64 university summer-school students.)

Variables	Vocabulary	Intelligence	Age	Semesters in college
(1)	(2)	(3)	(4)	(5)
Vocabulary	—	.40	.46	.21
Intelligence	.40	—	.29	.40
Age	.46	.29	—	.15
Semesters in college	.21	.40	.15	—

1. How great are the chances that age really bears a closer relation to vocabulary than intelligence does?
2. What are the chances that one's intelligence score could be predicted more accurately from his vocabulary than from his age?
3. What would be the correlation between intelligence and vocabulary if age were held constant? What are the chances that there is a real value in holding the age factor constant?
4. What would be the highest correlation possible if age and intelligence were combined in optimum manner and correlated with vocabulary? Would the difference be high enough to warrant making the combination? What are the chances?

CHAPTER XII

PRINCIPLES OF STATISTICAL RESEARCH

We have now described and shown how to use a few of the most common statistical tools that are ordinarily used in such research activities as confront the teacher and school administrator. It remains to indicate a few of the general principles and viewpoints which should govern the statistical work of persons in the field of educational research. They are as follows:

Knowledge of the field in which the investigation lies. Success in statistical research in education demands that the research worker be informed about education, as well as about statistics. If the investigation has to do with the teaching of chemistry he must know something about chemistry. One may be a thorough master of statistical techniques, and yet make serious blunders because of his ignorance of the field in which he is applying his statistics.

Purposeful work. Statistical work should not be treated as an end in itself. It should not be a form of "intellectual whittling." Merely figuring a number of correlation coefficients or drawing a number of beautiful graphs does not constitute statistical research. There is a tendency for beginners who have recently discovered how to do a few statistical processes to practice their recently acquired skills in very much the same way that the small boy with his first pocket knife indulges in whittling. Furthermore,

there is a tendency on the part of a small percentage of statistical students to become obsessed with the fancy-work idea, and to spend an unnecessary amount of time and energy on the purely physical task of drawing beautiful graphs or preparing tables of data so that they will look attractive to the eye. Neatness in statistical work is a virtue, of course, but it should not become an obsession to the extent of obscuring the main purpose, which is the solution of worth-while problems.

Previous planning of procedure. A statistical investigation should not be allowed to "grow up," as Topsy did, but should be carefully planned beforehand so that the procedure will yield the kind of data that will warrant careful treatment. More attention devoted to the logic of the study before the data are collected will greatly improve the final result. It is often advisable to rehearse the study in a preliminary way with a small group, in order to perfect the procedure before going into it with large numbers.

Soundness of the basic data. It is very important that the original data of an investigation be basically worth treating by refined statistical procedures. The average of a thousand guesses is not necessarily a fact and the policy of basing elaborate calculations upon data which in the beginning represented merely guesses is open to the severest criticism.

Adequate number of cases. Not only should the data be sound, but there should be enough cases to warrant generalization. The reliability of statistical conclusions is very definitely dependent upon the number of cases, and

the ever-present danger in statistical work is that of generalizing upon too limited a number of observations or measurements. Just as a building which rests on a very weak foundation cannot stand, so a statistical study that is based on only a few cases is certain to bring discredit upon the science of education. Furthermore, the policy of making an extra number of statistical computations as a means of compensating for the lack of cases is questionable, since it really amounts to building a more elaborate superstructure on an edifice in order to take attention away from the shaky foundation.

Representativeness of data. The cases that are observed, measured, or recorded as the basis for statistical treatment should be typical of the situation which they are supposed to represent. If conclusions are desired regarding methods of teaching elementary-school children, an experiment based upon white rats or upon feeble-minded children in a special institution would be questionable because of the lack of representativeness. Either the cases must be chosen for study with more consideration of the total field to which the conclusions are to apply, or else the conclusions which are drawn must be very definitely restricted in their application to the particular situation in which the study was carried out.

Many-sided analysis. A high-grade statistical investigation should involve data of many kinds, as well as data regarding a large number of cases. An effort should be made to get a variety of types of information, so that the problem can be viewed from more angles or slants. One

should get the whole body of facts regarding the situation, instead of merely scratching the surface. Just as many of the potentialities for getting the whole truth regarding a case in court may be thrown away because of not having the proper witnesses, or because of excluding certain testimony upon legal grounds, so many statistical investigations fail to get at the heart of the matter because of the lack of necessary data or the failure to use effectively everything that was known about the situation studied.

Records. An adequate system of keeping records for the study is very essential. The records should include entries regarding the various accompanying circumstances, so that at a later time it will be possible to check back to the original papers and find whether some important factor that has recently come to attention was adequately accounted for, or whether some important causal element was satisfactorily controlled. The plan of collecting all documents related to a given study, and filing them away for a considerable period of time after the study is completed, has much in its favor because there will be occasions when the opportunity to go back to the sources will prove invaluable.

Intelligent *versus* mechanical use of statistical tools. The person who knows only enough about statistics to be able to juggle formulas blindly, but does not know what purpose they serve nor what significance attaches to the results of their use, is likely to accomplish very little. One should study the processes and tools of statistics so that he knows why and how he uses them, and should not be satisfied with

merely acquiring the ability to get the mathematical answer. He should be able to state, in non-statistical terms intelligible to the ordinary layman, the main things which his statistical processes reveal. Unless the work is meaningful enough that it can be so stated in common-sense phraseology there is need for further study of the basic theory of the processes used.

Emphasis upon principles and relationships rather than upon facts. Statistics in the sense of numerical facts justly deserves to be called dry. A statistical study should do more than merely state facts. It should not stop with merely describing, but should go on to the more advanced stage of explaining, generalizing, showing causes and results, or pointing out relationships which exist between facts. The descriptive function is really the beginning point rather than the end, because the real work of the statistician is to discover trends, tendencies, relationships, and other hidden secrets which are contained in the body of dry facts used as the original data for the investigation.

Graphic methods to supplement numerical ones. The use of graphs to represent statistical quantities is advantageous not only because it helps to show up the essential findings in a way which others can appreciate and understand more clearly, but also because it serves as a splendid check upon the accuracy and soundness of the numerical work, and stimulates the mind of the investigator to analyze and discover more of the significant aspects of his data than he would detect if he relied purely upon the numbers themselves.

The rational element. Occasionally one reads a statistical report which suffers from excessive objectivity. By this we mean a report which omits the element of reasoning or critical analysis of the statistical findings. Figures do not interpret themselves. One must use his common sense or critical reasoning faculties to carry the work on beyond the point at which the mathematical work stops. After doing a thorough job in the strictly statistical part of the work one must finish his undertaking by doing a good job in the philosophical, rational, or subjective portion of the task. In other words, he must do some good thinking before, during, and after making mathematical calculations.

Validity of basic assumptions. Whether or not the data describe what they are assumed to describe, and whether or not the statistical processes employed really perform the functions which they are assumed to perform, are considerations of utmost importance. Occasionally one reads a report of an investigation which later proves to be utterly worthless because some one demonstrates the falsity of some of the basic assumptions. Thus Dr. Burt, the London psychologist, "thought he had proved statistically that a child's mental age is attributable to training much more than to endowment. Holzinger and Freeman show that by applying Burt's statistical methods to his own data it is possible to prove that a child's *chronological age* is also attributable to training; in other words, Burt's alleged proof is statistical nonsense." [1] When such blunders as this can be

[1] "The Possibilities and Limitations of Training," editorial in the *Journal of Educational Research*, vol. XIII, p. 371. (May, 1926.)

made by leading authorities in education and psychology, it is well for the beginner to be particularly cautious regarding what he takes for granted.

Criticism and supervision. Even the most careful statistician is likely to make mistakes, draw unwarranted conclusions, and overlook important aspects of his problem occasionally. The beginner is more likely to do so. Consequently a statistical study should be presented for the critical evaluation of others in order that these flaws may be detected and repaired. It is well to make contact with some one who is able to advise, guide, direct, and criticize, and it is well whenever possible to present the results of a study before a group of sympathetic and yet critical students of education in order that they may pick it to pieces and point out the ways in which it needs revision.

Checking for accuracy. Statistical work often is worse than useless if it contains mathematical errors which misrepresent facts. The very best statisticians make numbers of errors in their work, and they are so thoroughly aware of this fact that they generally have their work carefully checked before they leave it. This point has been mentioned before at various places in the chapters of this book, but it cannot receive too much emphasis.

Freedom from spurious accuracy. There is a phase of accuracy which is often overdone by amateurs in the field of statistics. This is illustrated by the carrying out of calculations to four decimal places, when the original data were of a very crude and inaccurate sort which would not warrant extreme refinement of calculations. If ten people guessed

at the height of a flag pole it would not be in good taste to average their guesses out to thousandths of an inch, yet this very thing is repeatedly done in educational statistics.

Computation of reliability. An essential of careful statistical work is that any generalization which is made on the basis of observed data should be accompanied by a statement of the reliability of that generalization. To say that one method of teaching is five points better than another method, or that it is 20 per cent better, is to make a very crude sort of statement. The careful statistician would probably say that, on the basis of the observed sampling of cases, the chances are 165 to 1, for example, that method A would prove to be better than method B, if the comparison were made for an infinite number of cases. Much of the early statistical work in education was of doubtful value because of the failure to compute reliability.

Alertness to discover hidden variables. The careful worker takes pains to see that when he is dealing with one or two variables he is not in reality dealing with a third one which works through one of these unawares. He is always suspicious that some other element may be lurking in the background and exerting an unsuspected influence on the situation being studied. He takes special pains to see that things are really comparable before he compares them. For example, he is on guard against making comparisons between totals when he should compare percentages, since totals for various items may be increasing at a time when percentages for the same items are on the decrease.

Critical use of the normal-curve assumption. Two com-

mon mistakes are made by beginners regarding the use of the normal curve in their statistical thinking. One is the mistake of assuming that all data conform to the normal curve, and the consequent treatment of all data by methods which are appropriate only to normally distributed data. The second error is that of assuming that because one's data conform to a normal distribution they are sound and valid. Thus the author of a questionnaire justified the soundness of his procedure and the value of his questionnaire on the ground that the replies, when tabulated, conformed to the normal curve. He apparently overlooked the fact that falsehoods and errors, as well as true statements, distribute themselves according to the normal curve.

Distinguishing between correlation and causation. The fact that two things are associated with each other does not prove that one is the cause of the other, nor does the fact that a variation in one is accompanied by a variation in the other indicate that the two are causally related. This tendency to mistake association or correlation for a causal relationship is exceedingly common.

Recognition of limitations. It is seldom that any statistical study will answer all the questions which we should like to have answered regarding a given problem. From the very nature of statistical work the results are likely to be partial and fragmentary, rather than complete and final. The statistician must make up his mind, therefore, to leave some questions unanswered and to admit freely that his study has limitations. He should even go beyond this point and call attention to the major shortcomings, limitations,

and deficiencies in his work, and warn his readers against the danger of attributing more to his investigation than is really claimed for it.

Freedom from prejudice. Statistical processes should not be used merely to justify prejudices nor to lend authoritativeness to rationalizations. Neither should the element of personal prejudice or bias enter into the process of interpreting the meaning of statistical results. It is not uncommon for amateurs to disregard their statistical findings almost completely when they come to draw conclusions, and to state dogmatically the prejudices with which they began, even when the statistical findings would seem to be directly opposite to those convictions.

Honesty. Some one has said that figures can't lie but that liars can figure. It should be unnecessary to emphasize the necessity of absolute integrity in dealing with statistical facts, and yet occasionally one sees efforts made deliberately to distort the truth in order to gain a point. Such a use of statistics in the service of personal ends is a more serious moral lapse than is the misrepresentation of facts in non-statistical situations, because it is a more cowardly thing to do. Statistical evidence carries with it an authoritativeness which makes it exceedingly harmful when used in the cause of falsehood, because statistically supported misrepresentations are more likely to be believed and accepted in good faith by those to whom they are made. It is the statistician who uses figures deliberately to deceive who is referred to in the statement that there are three kinds of liars, namely, "liars, infernal liars, and statisticians."

Coöperation with other research workers. In surveying the research that has been done on any particular problem, the impression which is commonly gained is that most of the work has been done in a very isolated and fragmentary way. Fifty or seventy-five research reports may have been published bearing upon a given problem, and yet apparently each study was made without much information about what the others had done or much effort to relate the new work to the old. Many times several people are pursuing research on a given problem, each being unaware of the fact that the others are attacking it. More effort should be made to get acquainted with what others have done and are doing, and to modify one's own procedure in such a way as to coördinate the work with that of other investigators.

Studying statistics in service. What the average beginning student learns about statistics in a course in the subject is usually very small and insignificant in comparison with what he learns when he begins to do a statistical investigation of his own. One should continue to learn statistics after entering the field of teaching, administration, or educational research, just as he should continue his study of educational psychology and related branches after he secures a teaching position. The study of statistical processes is more fruitful after one enters the practical field of research and educational service, because the new statistical techniques are learned in application to practical and concrete situations.

Recognition of the relation of statistics to the larger field of research and scientific method. It is a mistake to think

of statistics as synonymous with research and scientific method. It is a very important aspect of these, but not the only aspect. There are numerous research techniques which do not use statistics at all, or use it in a very limited way.[1] The scientific method is a type of thinking which has universal applications. Statistics is a subdivision of this larger field, and is comparable in rank to such other scientific procedures as the experiment, the historical study, the case study, the job analysis, the survey, etc. The student who has mastered the rudiments of statistics should not be content to let his achievements stop at that point, but should go on in quest of other tools for the pursuit of knowledge in order to be able to make the greatest possible contribution to the advancement of science.

TRUE-FALSE TEST

Decide whether each of the following statements is true or false, and be prepared to defend your decision.

1. There is no such thing as philosophical research.
2. Thorough training in methods of historical research would make one a better statistician.
3. Statistical workers have made too much of the need for a large number of cases.
4. One should never conclude that what is true of an observed group of cases is true for all.
5. The original data for an investigation are worthless as soon as the calculations are completed.
6. Most investigations are limited in their value because of uncontrolled variables.

---

[1] Some of these are discussed in Crawford, C. C., *The Technique of Research in Education.* University of Southern California, Los Angeles, 1928.

7. Mistakes in arithmetic are unimportant in comparison with chance errors due to sampling.
8. Statistical research should begin and end in philosophy.
9. Careful statistical treatment can usually compensate for inadequate or even erroneous data.
10. To drop decimals or make approximate calculations is a sure sign of a slovenly mind.
11. One should never do research on a problem upon which he is not already perfectly informed.
12. Statistical research involves a large amount of purely routine labor.
13. The best way to do a good piece of research work is to start on a problem and work at it until it expands into an investigation of major proportions.
14. Data should always be assumed to follow the normal curve unless the contrary is clearly proven.
15. A statistician who has a personal prejudice regarding a problem is mentally and morally unfitted for research on that problem.
16. Dishonesty has caused more erroneous conclusions in statistical research than has ignorance.
17. The value of graphs lies mainly in their helping to convey ideas to others.
18. Description is not the end but rather the beginning of statistical research.
19. A statistical study should never be made with less than 100 cases.
20. Perfect representativeness of data cannot be secured unless the entire population is included in the study.

QUESTIONS AND EXERCISES

1. What would you say is the greatest difficulty involved in making statistical investigations?
2. Illustrate cases of faulty conclusions you have known to be drawn from statistics, and tell how they came to be drawn.
3. How can the results of statistical research be made to affect educational practice to a greater degree?
4. Cite pieces of statistical research which are well known or exceptionally good, and tell what you consider the special points of merit in each.
5. Make up a score-card consisting of the five or ten major qualities or attributes of good statistical research, by which you would judge the work of an investigator, giving each item the weight you think its importance deserves. Justify your weighting of the items.

SELECTED REFERENCES

Almack, J. C. *Research and Thesis Writing.* Boston, Houghton Mifflin Company, 1930.

Brinton, W. C. *Graphic Methods of Presenting Facts.* New York, The Engineering Magazine Company, 1914.

Chaddock, R. E. *Principles and Methods of Statistics.* Boston, Houghton Mifflin Company, 1925.

Crawford, C. C. *The Technique of Research in Education.* Los Angeles, The University of Southern California, 1928.

Garrett, H. E. *Statistics in Psychology and Education.* New York, Longmans, Green, and Company, 1926.

Good, C. V. *How to Do Research in Education.* Baltimore, Warwick and York, Inc., 1928.

**Holzinger, K. J. *Statistical Methods for Students in Education.* Boston, Ginn and Company, 1928.

Holzinger, K. J. *Statistical Tables for Students in Education and Psychology.* Chicago, University of Chicago Press, 1925.

**Kelley, T. L. *Statistical Method.* New York, The Macmillan Company, 1923.

McCall, W. A. *How to Experiment in Education.* New York, The Macmillan Company, 1923.

McCall, W. A. *How to Measure in Education.* New York, The Macmillan Company, 1922.

Odell, C. W. *Educational Statistics.* New York, The Century Company, 1925.

Otis, A. S. *Statistical Method in Educational Measurement.* Yonkers-on-Hudson, World Book Company, 1925.

Pearson, K. *Tables for Statisticians and Biometricians.* Cambridge (England) University Press, 1924.

*Rugg, H. O. *A Primer of Graphics and Statistics.* Boston, Houghton Mifflin Company, 1925.

Rugg, H. O. *Statistical Methods Applied to Education.* Boston, Houghton Mifflin Company, 1917.

**Advanced. *Very elementary.

Thurstone, L. L. *Fundamentals of Statistics.* New York, The Macmillan Company, 1925.

*Whitney, F. L. *Statistics for Beginners in Education.* New York, D. Appleton Company, 1929.

*Williams, J. H. *Elementary Statistics.* Boston, D. C. Heath and Company, 1929.

Williams, J. H. *Graphic Methods in Education.* Boston, Houghton Mifflin Company, 1924.

**Yule, G. U. *An Introduction to the Theory of Statistics.* London, Charles Griffin and Company, 1911.

---

* Very elementary. ** Advanced.

INDEX

Accuracy, with slide rule, 19; of frequency tabulations, 30; of assumed mean, 56; of measurement, and reliability, 130; in research, 197; spurious, 197
Adding machine, computing percentiles on, 82
Analysis, in research, 193, 194
Assumed mean, 53–58, 98
Authenticity, and reliability, 130
Average, 49
Average deviation, 100–02
Ayres Handwriting Scale, scores on, 35

Barlow, *Tables*, 17
Bimodal distribution, 70, 125, 126
Binomial theorem, and normal curve, 109
Biology, and normal curve, 111
Buckingham, B. R., "Multiple Correlation," 181 *n.*
Burroughs calculating machine, 21
Burt, mental age, 196

Calculating machines, 20, 21
Causal relationship, not signified by correlation, 154, 199
Central tendency, mean as measure of, 49, 50; median as measure of, 62
Chance, and normal curve, 106–08; reliability application of, 131, 132; P.E. table to determine, 147–49
Class interval, 31–36; rule for finding size of, 32; midpoint of, 33, 34; graphic illustration of, 33; designation of, 36; tops of, cumulative frequencies converted into percentile ranks, 77–79; midpoints of, percentile ranks, 79–81
Coefficients of correlation, 143, 155; mathematical computation of, 158–62; Pearson Product-Moment Method of computing, 159, 160; interpretation of, 162–64; reliability of, 170; 174–80; data for practice in computing, 172; steps in finding partial, 181; multiple, 181–85. *See also* Correlation
Coin, tossing and normal curve, 107, 108; and reliability, 131; and standard error, 133–35
Comptometer, 21
Continuous, and discrete data, 36, 37
Coöperation, in research, 201
Correlation, 153–88; common use of term, 153; definition of, 153, 154; not necessarily causal relationship, 154; perfect positive to perfect negative, 155, 156; coefficients of, 155; revealed by scatter diagram, 156–58; mathematical computation of coefficient, 158–62; Pearson Product-Moment Method of Computing, 159, 160; interpretation of coefficient, 162–64; computed without scatter diagram and correlation table, 164; Pearson method for ungrouped pairs, 164–66; Spearman rank dif-

ferences method, 166–69; reliability of coefficient, 143; 170; 174–80; reliability and size of coefficient, 175; data for practice in computing coefficients, 172; number cases and reliability of coefficient, 175; formula for standard error of coefficient, 176; increasing reliability of coefficient, 177; reliability of difference between two coefficients, 177–79; P.E. as measure of reliability of coefficients, 179, 180; partial, 180, 181; multiple, 181–85; P.E. and partial and multiple, 185, 186; curvilinear, 186–88; data for practice with special aspects of, 190; and causation, 199

Correlation work-sheets, 24

Crawford, C. C., *Technique of Research in Education*, 202 *n.*

Cumulative-frequency table, as aid in computing percentiles, 76, 77; converting into percentile ranks, 77–79

Curves, and frequency polygons, 39–41

Curvilinear correlation, 186–88

Dalton calculating machine, 21

Data, tabulation of, 28–45, 30, 31; continuous and discrete, 36, 37; normal and skewed, 68; ungrouped, percentile ranks computed from, 86–88; representativeness of, 193

Data sheets, as labor-saving devices, 23, 24

Decile, relation of percentile to, 74; to describe variability, 95

Decimals, in using tables, 15

Deviation, quartile, 95. *See also* Quartile deviation; standard, 96–100; average or mean, 100–02

Difference, observed and true, 139; true, determining chances that it is above zero, 140, 141

Discrete, and continuous data, 36, 37

Dispersion. *See* Variability

Distributions, "bimodal," 70; two, compared by percentile graph, 85, 86; two, variability in comparing, 93; normal, *see* Normal curve

Education, statistics in study and practice of, 1–9; statistical research in, 191–202

Educational situations, illustrating use of percentiles, 75, 76

Error, statistical training protection against, 3

Examinations, normal curve to determine difficulty of, 119, 120; difficulty-value of questions, 122, 123

Foreign language, statistics substituted for in Ph.D. requirement, 8, 9

Freeman, F. N., criticism of Burt, 196

Frequencies, percentage, in comparing groups of unequal size, 43–45

Frequency distribution, nature of, 28–30; illustration of with tabulation, 29; grouped in class intervals, 31; graphic representation of, 38, 40; in comparing groups of unequal size, 42, 43

Frequency polygons, and curves, 39–41; comparison of groups by, 41, 42; to represent variability, 94

Frequency rectangles, 37–39

Frequency tables, finding mean from, 51–57; finding median from, 63–66
Frequency tabulation, 29

Grading, uniform, by normal curve, 117–19. *See* Marks
Graph, percentile, 82–84
Graphic methods, in research, 195
Graphic representation, of frequency distributions, 38, 40
Groups, comparison of, by frequency polygons, 41, 42; of unequal size compared by simple frequencies, 42, 43; of unequal size, percentage frequencies in comparing, 43–45; two, compared by percentile graphs, 86

Haggerty Reading Examination, tabulation of scores in, 41
Histograms, 37–39
Hollerith machine, 21–23
Holzinger, K. J., *Statistical Tables for Students in Education and Psychology*, 16; correlation work-sheet, 24; *Statistical Methods for Students in Education*, 146 *n.*; criticism of Burt, 196
Honesty, in research, 200

Interpolation, in using tables, 13–16
I.Q., tabulation of, 31

Kelley, T. L., correlation work-sheet, 24; *Statistical Method*, 99 *n.*; 134 *n.*, 165 *n.*

Labor-saving devices, 12–24; importance of, 12, 13; tables, 13–18; slide rule, 18–20; calculating machines, 20–21; Hollerith machine, 21–23; printed forms and data sheets, 23, 24
Lefax Company, The, *Tables*, 17
Logarithmic tables, 17, 18

Marchant calculating machine, 21
Marks, school, use of statistics in assigning, 4, 5; and normal curve, 117–19
Mathematics, normal curve and, 109, 110
McCall, *How to Experiment in Education*, 137, 139 *n.*, 142, 181 *n.*
Mean, 49–59; and "average," 49; measure of central tendency, 49, 50; usual method of computing, 50, 51; finding from frequency tables, 51–57; assumed, 53–58; assumed, testing accuracy of computation from, 56; computed from grouped and ungrouped data compared, 58, 59; definition of, 59; median compared with, 62, 63; and median, where to use, 68, 69; and standard deviation, 114, 115; reliability of, 133–35; standard error of, 135; determining above given point, 136; reliability of difference between two, 138–40; reliability of sum of two or more, 141; reliability of difference between sums of two sets, 141
Mean deviation, 100–02
Median, 62–70; definition of and comparison with mean, 62, 63; finding from frequency table, 63–66; of ungrouped data, 66, 67; and mean, when to use, 68, 69; relation of percentiles to, 74; from percentile graph, 84; reliability of, 142

Midpoint, of class interval, 33, 34
Mode, 69, 70
Monroe calculating machines, 21
Multiple correlation, 181–85

Normal curve, 106–26; and chance, 106–08; and mathematics, 109, 110; approximation of educational and biological data to, 110, 111; further mathematical properties of, 111, 112; area and standard deviation, 112–15; table showing percentages of, between mean and σ unit values, 115; and distribution of teachers' marks, 117–19; to determine difficulty of examinations, 119, 120; and product scale, 123, 124; significance of hypothesis of, 124, 125; critical use of in research, 198, 199

Otis, correlation work-sheet, 24; "Universal Percentile Graph," 24

Partial correlation, 180, 181
Pearson Product-Moment Method, of computing coefficient of correlation, 159, 160; for ungrouped pairs, 165
Percentage frequencies, 43–45
Percentile curve, to describe variability, 94
Percentile graph, 82–84; to compare two distributions, 85, 86
Percentile ranks, definition of, 74, 75; converting cumulative frequencies into, 77–79; of midpoints of class intervals, 79–81; computed from ungrouped data, 86–88
Percentiles, 73–88; meaning of, 73, 74; relation to median, quartiles, quintiles, and deciles, 74; and percentile ranks, definition of, 74, 75; educational situations using, 75, 76; cumulative-frequency table as aid in computing, 76, 77; short methods of computing, 81, 82; from percentile graph, 84
Prejudice, freedom from, in research, 200
Printed forms, as labor-saving devices, 23, 24
Probability, reliability and, 131. *See also* Chance
Probable error, to measure reliability, 133, 144; and standard error, as measures of reliability, 144–46; for measuring reliability of correlation coefficients, 179, 180; and partial and multiple correlation, 185, 186
Probable error table, to determine chances, 147–49
Product scale, 123, 124

Quartile, relation of percentile to, 74; from percentile graph, 84; and quartile deviation, 95
Quartile deviation, 95; calculation of, 96–100
Quintile, relation of percentile to, 74; to describe variability, 95

Range, of scores in class interval, 32; as measure of variability, 94
Rank-differences, 166–69
Ranks, percentile, definition of, 74, 75
Records, in research, 194
Reliability, 129–50; statistical, 129; what it does not include, 129–31; application of theory of chance, 131, 132; dependent on variability,

132; dependent on observed number of cases, 132; standard error unit for measurement of, 133; number cases to produce given degree of, 136–38; of difference between two means, 138–140; of sum of two or more means, 141; of difference between sums of two sets of means, 141; in typical rotation experiment, 142; of other statistical quantities, 142; probable error and standard error as measures of, 144–46; formulas for using P.E. instead of, 146, 147; of correlation coefficients, 170; 174–80; of correlation coefficients, probable error as measure of, 179, 180; of coefficient of correlation, increase in, 177; of difference between two coefficients of correlation, 177–79; computation of in research, 198

Representativeness, and reliability, 130

Research, statistics and, 6; principles of statistical, 191–202; importance of knowledge of field, 191; purposeful work in, 191, 192; planning, 192; adequate number of cases, 192, 193; soundness of basic data, 192; representativeness of data, 193; many-sided analysis, 193, 194; keeping records, 194; intelligent use of tools, 194, 195; facts *vs.* principles and relationships, 195; graphic methods, 195; the rational element, 196; validity of basic assumptions, 196; criticism and supervision, 197; spurious accuracy, 197; checking for accuracy, 197; computation of reliability, 198; discovering hidden variables, 198; critical use of normal curve, 198, 199; correlation and causation, 199; recognition of limitations, 199; freedom from prejudice, 200; honesty in, 200; coöperation in, 201

Rotation experiment, reliability in, 142

Rugg, H. O., *Statistical Methods Applied to Education*, 163 *n.*, 167 *n.*

Scatter, 91

Scatter diagram, correlation revealed by, 156–58

Scientific method, relation of statistics to, 201, 202

Score, single, meaning of, 34, 35

Slide rule, 18–20; computing percentiles on, 82

Social sciences, statistics and, 7–9

Spearman, rank differences method. 166–69

Standard deviation, 96–100; computation of illustrated, 99, 101; and area of normal curve, 112–15; table of, 115; of mean, 133–35; for measurement of reliability, 135; unit values of, and chances that measures will not fall below, 137; computation of reliability of, 143; and probable error, as measures of reliability, 144

Statistical research, principles of, 191–202. *See* Research

Statistical tools, intelligent use of, 194, 195

Statistics, educational importance of, 1–9; teachers' need for, 1, 2; uses for in study of education, 2, 3; as protection against error, 3; need

for in conduct of educational activities, 3–5; a method of thinking, 5; in improvement of education, 6, 7; and the social sciences, 7–9; substituted for foreign language in Ph.D. requirement, 8, 9; studying in service, 201; relation to scientific method, 201, 202

Tables, as labor-saving devices in statistics, 13–18; interpolation in use of, 13–16; decimals in use of, 15; books of, 16, 17; logarithmic, 17, 18
Tabulation, of data, 28-45
Teachers' marks, and normal curve, 117–19
Thinking, statistics a method of, 5
Thurstone, L. L., correlation worksheet, 24; *Fundamentals of Statistics*, 109
Toops, correlation work-sheet, 24

Ungrouped data, finding mean from, 15, 58; finding median of, 66, 67

Validity, and reliability, 130; of basic assumptions in research, 196
Variability, 91–103; meaning of, 91; need for measures of, 91–94; in comparing two distributions, 93; methods of measuring, 94; range as measure of, 94; frequency polygon to represent, 94; described by percentile curve, 94; quintiles, deciles, to describe, 95; quartiles to describe, 95; standard deviation, 96–100; computation of two principal measures of, 100; average or mean deviation, 100–02; when to use each measure of, 102, 103; reliability dependent on, 132; σ as measure of, 133
Variables, hidden, in research, 198

Yule, G. U., *Introduction to Theory of Statistics*, 109 *n.*